Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2021.018577
Article

Tech Science Press

A Shadowed Rough-fuzzy Clustering Algorithm Based on Mahalanobis Distance for Intrusion Detection

Lina Wang[1,2,*], Jie Wang[3], Yongjun Ren[4], Zimeng Xing[1], Tao Li[1] and Jinyue Xia[5]

[1]School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[2]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, 519080, China
[3]School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[4]School of Computer and software, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[5]International Business Machines Corporation (IBM), New York, NY, 10504, USA
[*]Corresponding Author: Lina Wang. Email: wangln@nuist.edu.cn
Received: 12 March 2021; Accepted: 23 April 2021

Abstract: Intrusion detection has been widely used in many application domains; thus, it has caught significant attention in academic fields these years. Assembled with more and more sub-systems, the network is more vulnerable to multiple attacks aiming at the network security. Compared with the other issues such as complex environment and resources-constrained devices, network security has been the biggest challenge for Internet construction. To deal with this problem, a fundamental measure for safeguarding network security is to select an intrusion detection algorithm. As is known, it is less effective to determine the abnormal behavior as an intrusion and learn the entire scope of the normal behavior with the traditional anomaly-based algorithm for Internet intrusion detection. In this paper, we propose an intrusion-detecting algorithm of shadowed rough-fuzzy clustering based on Mahalanobis distance, named MSRFCM. It adopts dissimilarity measurement of Mahalanobis distance to identify the relevant variables that significantly influence the clustering performance and reduce the error rate in the process of partitioning clusters with high attribute correlation. And shadowed rough-fuzzy clustering (SRFCM) is applied to obtaining real value-approaching prototypes based on iteration and partitioning the data set into more meaningful clusters. Through simulation with the NSL-KDD intrusion data set and three other intrusion data sets, the Mahalanobis distance-based shadowed rough-fuzzy clustering algorithm has improved performance in intrusion detection.

Keywords: Intrusion detection; security; SRFCM; FCM; Mahalanobis distance

1 Introduction

As network technologies develop, the network environment becomes more complex, and more systems are vulnerable to intrusion attacks. With the rapid development of Internet environments, there is an increasing demand for securing all kinds of internet environments [1–5]. Data mining algorithms have

recently been applied to network intrusion detection [6–8]. Data mining-based network intrusion detection technologies differ from others due to their data-centric models, regarding intrusion detection as a procedure for analyzing and processing massive security audit records. The clustering technique plays an important role in data analysis and interpretation as it gains insight into the nature of the data pattern by discovering hidden structures in it. Fuzzy *c*-means clustering algorithm (FCM) is widely used in intrusion detection, and it categorizes similar samples (patterns) into clusters, but it is an approach somewhat unsatisfactory for intrusion detection processing due to its inconsistent outcomes and liability to a local minimum, and subsequently leading to low detection rate [9–11]. In view of these restrictions and its sensitiveness to the existence of noisy data, a rough fuzzy clustering algorithm (RFCM) is proposed, incorporating fuzzy sets to capture uncertainty associated with the samples [12,13]. RFCM methods, however, need to optimize two sets of parameters to achieve the best clustering objective best at each iterative step. The experiment has been implemented using GAs to tune the weighted coefficient and threshold parameter by minimizing a fitness function based on certain clustering validity indices [14]. In order to disambiguate and capture the essence of a distribution, the concept of shadowed sets has been introduced in the literature, such as work by Pedrycz [15,16].To ensure a more logical selection of threshold parameters, a shadowed rough-fuzzy *c*-means clustering (SRFCM) was proposed to automatically select a threshold parameter [17], where all clustered patterns are placed into three categories: complete belongingness (core level), complete exclusion (exclusion level), and unknown (boundary level), assuming a particular perspective built by an optimization process. A series of rough fuzzy clustering algorithms based on shadowed sets have been proposed by some scholars [18–20].

The definitions of dissimilarity measures can also improve the clustering algorithm. All of the above algorithms apply the Euclidean distance to measure the dissimilarity between two samples without considering the difference of each attribute's relative importance. Mahalanobis distance [21] is introduced to the algorithm that neither sample attributes coupling nor actual dimension is taken into consideration in the cluster partitioning process. Since it accounts for unequal variance as well as correlations between attributes, it adequately evaluates the distance by assigning different weights or import factors to the attributes of samples. In addition, Mahalanobis distance can be used to readjust the geometric distribution of patterns so as to reduce the distance of similar patterns, which can prevent increasing the error rate of Euclidean distance in calculating high attribute correlation of data sets, when the sample distribution follows Gauss distribution [22]. Subsequently, a Mahalanobis distance-based fuzzy clustering algorithm (MFCM) is proposed, by using Mahalanobis distance instead of Euclidean distance in traditional FCM clustering [23,24]. The MFCM algorithm's accuracy increases obviously when dealing with the data sets with high attribute correlation and can effectively resolve the deficiency with FCM to induce aspheric clusters. It will be an issue of singularity when the calculation of Mahalanobis distance involves the inversion of the covariance matrix. In this case, eigenvalue, eigenvector and pseudo-inverse operations are utilized to deal with it.

To improve intrusion detection performance, a shadowed rough-fuzzy clustering intrusion detection algorithm based on Mahalanobis distance (MSRFCM) is proposed. In this research, the SRFCM algorithm is applied to obtain real value-approaching prototypes at iteration and partition the data set into more meaningful clusters. Besides, the dissimilarity measurement of Mahalanobis distance is utilized to identify the relevant variables that have a significant influence on clustering performance and reduce the error rate in the partitioning cluster process with high attribute correlation. The feasible solution for better clustering results is provided by integrating Mahalanobis distance with shadowed rough-fuzzy clustering. Four intrusion data sets known as NSL-KDD, AWID, UNSW-NB 15, and CICIDS-2017, are employed for experimental purposes regarding intrusion detection in computer networks [25]. Experiments on these data sets demonstrate that MSRFCM, the novel algorithm, has improved the intrusion detection performance using validity indices including Recall, Precision, and F1 score.

Organization of the paper is as follows. Section 2 provides the shadowed rough-fuzzy clustering algorithm. Section 3 discusses a Mahalanobis distance-based dissimilarity measure. The proposed algorithm is elaborated in Section 4. Experimental analysis is made to validate the advantage of the algorithm in Section 5. Finally, Section 6 concludes the study and highlights potential future work.

2 Shadowed Rough-Fuzzy Clustering

In 1982, a Polish scholar Z. Pawlak put forward a rough set theory–a mathematical tool for describing incompleteness and uncertainty. It provides an effective approach to analyzing inconsistent, inaccurate, and other incomplete information and also functions in data analysis and reasoning to reveal rules hidden behind patterns. The rough set is characterized by its upper bound $\overline{B}X$and its lower bound $\underline{B}X$, which means the samples definitely belonging to a cluster occur within the lower bound and the samples that possibly belong to a cluster occur between the lower bound and the upper bound, namely the boundary region. The following basic properties of the rough set need to be satisfied.

Property 1: A sample can belong to the lower bound of one cluster at most.

Property 2: A sample that belongs to the lower bound of a cluster also belongs to the upper bound of the same cluster.

Property 3: A sample that does not belong to any lower bound belongs to more than one upper bounds.

The shadowed set is an improvement of fuzzy set through information simplification and retention of key fuzzy information. In shadowed sets, three quantification levels describing the elements of the set 0, 1, and [0,1], are utilized to simplify the fuzzy relation. Conceptually, shadowed sets are close to rough sets even though the mathematical foundations are very different. The concepts of the negative region, lower bound, and boundary region correspond to three logical values 0, 1, and [0,1] in shadowed sets, namely, excluded, included and uncertain, respectively. The unknown is formally termed shadowed region.

In shadowed sets theory, the threshold parameter is automatically obtained from fuzzy membership partition to determine the approximation regions for each cluster. The construction of shadowed sets is based on balancing the uncertainty that is inherently associated with fuzzy set. By elevating membership values of some regions of the universe to 1, and at the same time, reducing membership values of some regions of the universe to 0, we can eliminate the uncertainty in these regions. In order to balance the total uncertainty regions, it needs to compensate for these changes by allowing for the emergence of uncertain regions, namely shadowed sets. The main merits of shadowed sets involve the optimization mechanism for choosing the threshold and the burden reduction of the plain numeric computations.

Assuming X is a data set and u_{ij} is the probabilistic membership of pattern x_j to some cluster, where $u_{ij} \in [0,1]$. To obtain the optimal threshold, $u_{\max}$ and $u_{\min}$ are defined as the maximal and minimal membership of each cluster, respectively. And then, the range of feasible threshold values is $[u_{\min}, (u_{\min}+u_{\max})/2]$ [17]. Next, optimize the objective function γ_i is optimized according to the shadowed set in the i^{th} cluster, by following Eq. (1).

$$\gamma_i = \Big| \sum_{j:u_{ij}<\beta_i} u_{ij} + \sum_{j:u_{ij}>u_{\max}-\beta_i} (u_{\max}-u_{ij}) - card\{x_j|\beta_i \le u_{ij} \le u_{\max}-\beta_i\} \Big| \tag{1}$$

where $\sum_{j:u_{ij}<\beta_i} u_{ij}$ is the sum of membership for those patterns that are not the part of the cluster, $\sum_{j:u_{ij}>u_{\max}-\beta_i} (u_{\max}-u_{ij})$ is the sum of membership for patterns belonging to the cluster, and $card\{x_j|\beta_i \le u_{ij} \le u_{\max}-\beta_i\}$ represents the shadowed set. Then, the optimal threshold $\beta_i = \arg\min(\gamma_i)$ is determined for the i^{th} clusters.

Based on the obtained threshold β_i, the upper and lower approximation sets are expressed by Eqs. (2) and (3).

$$\bar{B}X_i = \{x_j | u_{ij} > \beta_i\} \tag{2}$$

$$\underline{B}Xi = \{x_j | u_{ij} > u_{max} - \beta_i\} \tag{3}$$

Further, the boundary region is calculated with Eq. (4).

$$BNP(X_i) = \bar{B}X_i - \underline{B}X_i = \{x_j | \beta_i \leq u_{ij} \leq u_{max} - \beta_i\} \tag{4}$$

Subsequently, the SRFCM algorithm is constructed to respond to the above discussions. Suppose these samples are classified as c and x_j to represent any test datum belonging to the i^{th} cluster with the membership degree u_{ij}. Thus, the objective function of the algorithm can be described with Eq. (5).

$$J(u_{ij}, v_i) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m'} ||x_j - v_i||^2 \tag{5}$$

where m' is the fuzzifier exponent with $m' = 2$. v_i is the prototype corresponding to the i^{th} cluster, $u_{ij} \in [0, 1]$ is the probabilistic membership of pattern x_j to some cluster, and $\|\cdot\|$ is the distance norm. Then the clustering aims to find the minimum of the objective function by iteration. Incorporating rough sets and shadowed sets with automatically obtained threshold parameters, the results of clusters partition achieve better. The prototype is calculated by Eq. (6).

$$v_i = \begin{cases} w_{\text{low}} \dfrac{\sum_{X_j \in \underline{B}X_i} u_{ij}^{m'} X_j}{\sum_{X_j \in \underline{B}X_i} u_{ij}^{m'}} + w_{\text{up}} \dfrac{\sum_{X_j \in (\bar{B}X_i - \underline{B}X_i)} u_{ij}^{m'} X_j}{\sum_{X_j \in (\bar{B}X_i - \underline{B}X_i)} u_{ij}^{m'}}, \underline{B}Xi \neq \varnothing \wedge \bar{B}X_i - \underline{B}X_i \neq \varnothing \\ \dfrac{\sum_{X_j \in (\bar{B}X_i - \underline{B}X_i)} u_{ij}^{m'} X_j}{\sum_{X_j \in (\bar{B}X_i - \underline{B}X_i)} u_{ij}^{m'}}, \underline{B}Xi = \varnothing \wedge \bar{B}X_i - \underline{B}X_i \neq \varnothing \\ \dfrac{\sum_{X_j \in \underline{B}X_i} u_{ij}^{m'} X_j}{\sum_{X_j \in \underline{B}X_i} u_{ij}^{m'}}, \text{other} \end{cases} \tag{6}$$

where $\underline{B}X_i$ and $\bar{B}X_i$ denote the lower and upper bounds of the cluster Xi, respectively. $\bar{B}X_i - \underline{B}X_i$ denotes the boundary region of the cluster X_i. $\dfrac{\sum_{X_j \in \underline{B}X_i} u_{ij}^{m'} X_j}{\sum_{X_j \in \underline{B}X_i} u_{ij}^{m'}}$ and $\dfrac{\sum_{X_j \in (\bar{B}X_i - \underline{B}X_i)} u_{ij}^{m'} X_j}{\sum_{X_j \in (\bar{B}X_i - \underline{B}X_i)} u_{ij}^{m'}}$ can be considered as the contributors to the fuzzy lower and fuzzy boundary regions separately. The coefficient w_{low}, as the weight of the lower bound samples, is crucial, whose value should range in [0.5,1] and $w_{\text{up}} = 1 - w_{\text{low}}$. The performance of the algorithm is dependent on the choice of w_{low}, w_{up}, and the threshold. Patterns in the lower bound significantly contribute to the prototype, and those patterns in the boundary region make a minor contribution to the prototype; thus, it is beneficial to obtain reasonable prototype sets and produce better clustering results.

The iteration of the fuzzy membership is denoted with Eq. (7).

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\dfrac{||x_j - v_i||}{||x_j - v_k||}\right)^{\left(\dfrac{2}{m' - 1}\right)}} \tag{7}$$

The specific steps of SRFCM algorithm are as follows.

Algorithm 1: SRFCM

Input: $X = \{x_1, x_2, \ldots, x_n\}$, the cluster number c, the fuzzifier exponent m', the lower approximate weight w_{low}, the iterative termination error ϵ, and the number of iterations t.

Output: All generated clusters and objective function values.

1: Random initialize the membership matrix U;

2: According to the shadowed set in the i^{th}cluster, optimize threshold β_i through objective function γ_i with Eq. (1);

3: Determine the upper and lower approximation sets by threshold β_i obtained in Step 2 with Eqs. (2) and (3), and calculate boundary region with Eq. (4);

4: Calculate the cluster prototype V with Eq. (6);

5: Calculate new membership degree with Eq. (7);

6: Repeat Step 2 to 5 until the termination condition is satisfied.

As the presence of approximated patterns between upper and lower bounds, the identification of normal or abnormal patterns in the boundary region is dilemmatic. For comparative analysis of a series of different algorithms, the maximal calculated value of membership degree decides the pattern to a certain cluster in SRFCM algorithm. Simulation experiments show that the definition w_{low} is also crucial besides the threshold parameter β, and w_{low} has a fixed value at each experiment. Attempts should be made in locating the best value of w_{low} for different data sets based on clustering validity indices such as DB, Dunn, and XB in practice.

3 Mahalanobis Distance

Calculating Mahalanobis distance involves the inverse of a covariance matrix Σ, which is often singular and leads to the inability to solve Mahalanobis distance directly. Both eigenvalue decomposition [26] and matrix inner product [27] are commonly used to resolve the issue. Here matrix inner product is applied to this research.

Define X as a sample matrix of $m \times l$, where m is the row number of a random vector $x_i, i = 1, 2, 3, \ldots, m$. Some statistical variables can be expressed in the form of sample matrices.

The sample mean vector v is calculated by Eq. (8).

$$v = X^T L \tag{8}$$

The sample covariance matrix C is expressed as Eq. (9).

$$C = \frac{1}{m} X^T X - X^T L L X \tag{9}$$

where L is a $m \times m$ matrix with each component equal to $\frac{1}{m}$.

The sample inner product matrix K is defined as Eq. (10).

$$K = \left\{x_i, x_j\right\}, i, j = XX^T \tag{10}$$

And the centered matrix K_c is expressed as Eq. (11).

$$K_c = K - LK - KL + LKL \tag{11}$$

where K and K_c are real symmetric semi-definite matrices.

Meanwhile, K_c can be decomposed with Eq. (12):

$$K_c = \alpha^T \Omega \alpha \tag{12}$$

where α is defined as the matrix composed of the eigenvector of K_c, and Ω is expressed as the diagonal matrix, whose diagonal element is comprised of the eigenvalues of K_c.

To obtain C^+, the pseudo-inverse matrix of covariance with Eq. (13) is utilized.

$$C^+ = mX^T \alpha^T \Omega^{-2} \alpha X \tag{13}$$

In Eq. (13), Ω^{-2} denotes the square pseudo inverse of Ω and C^+ can be calculated step by step from the inner product matrix of the sample in the input space.

If the sample is non-linear separable, then its non-linear mapping is employed into a high-dimensional feature space. To avoid explicitly defining the non-linear mapping, a kernel function can be used to replace the inner product in the feature space [28]. At this point, for the inner product matrix $K = \{K(x_i \cdot x_j)\}_{i,j}$, the Mahalanobis distance in feature space from Eq. (13) is expressed as Eq. (14).

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} = \sqrt{(Xx - X\mu)^T m \alpha^T \Omega^{-2} \alpha (Xx - X\mu)} \tag{14}$$

The empirical kernels of sample x_i and the mean μ on sample population X are mapped as:

$$Xx_i = K(X, x_i) = (K(x_1 \cdot x_i), K(x_2 \cdot x_i), \cdots, K(x_n \cdot x_i))^{\mathrm{T}} \tag{15}$$

$$X\mu = (K(x_1 \cdot \mu), K(x_2 \cdot \mu), \cdots K(x_n \cdot \mu))^{\mathrm{T}} \tag{16}$$

The distance can be calculated through the inner product expressed by the kernel function without involving any non-linear mapping. It should also be noted that the pseudo-inverse matrix of the covariance is related to the inner product matrix, which is no longer related to the dimension of the eigenvector but to the number of samples. Therefore, it brings computational advantages in high-dimensional space.

There are two operations to solve the inverse matrix of the covariance matrix. If the sample is linearly separable, an inner product matrix operation is applied directly. Otherwise, given $K(x_i, x_j) = e^{||x_i - x_j||^2/(2\sigma^2)}$, the radial basis function is selected as the kernel function with σ at 0.5 and is adopted in matrix inner product operation.

4 Mahalanobis Distance-Based Shadowed Rough-Fuzzy Clustering Algorithm

4.1 Mahalanobis Distance-based Fuzzy Clustering

On the basis of classical FCM objective function, Euclidean distance is replaced by Mahalanobis distance, and a covariance adjusting factor $-\ln|\sum_i^{-1}|$ is introduced to the objective function of the MFCM algorithm, which is defined as Eq. (17).

$$J_{\mathrm{MFCM}}(U, V, \Sigma, X) = \Sigma_{i=1}^{c} \Sigma_{j=1}^{n} \left[(x_j - v_i)^T \Sigma_i^{-1} (x_j - v_i) - \ln|\sum\nolimits_i^{-1}| \right] u_{ij}^{m'} \tag{17}$$

Its constraint condition is $\sum_{i=1}^{c} u_{ij} = 1, j = 1, 2, 3 \ldots, n,\ 0 \le u_{ij} \le 1$. The Lagrange operator for this optimization problem is solved by Eq. (18).

$$J = \Sigma_{i=1}^{c}\Sigma_{j=1}^{n}[(x_j - v_i)^T\Sigma_i{}^{-1}(x_j - v_i) - \ln|\sum\nolimits_i{}^{-1}|]u_{ij}^{m'} + \Sigma_{j=1}^{n}\alpha_j(1 - \Sigma_{i=1}^{c}u_{ij}),$$
$$0 \le \alpha_j \le 1, j = 1, 2, \ldots, n \tag{18}$$

Then, the operator is minimized, the partial derivative of v_i, α_j, u_{ij} to J is solved, and both of them are set as zero:

From $\dfrac{\partial J}{\partial v_i} = 0$,

$$v_i = \frac{\Sigma_{j=1}^{n}u_{ij}^{m'}x_j}{\Sigma_{j=1}^{n}u_{ij}^{m'}}, i = 1, 2, 3, \ldots, c \tag{19}$$

From $\dfrac{\partial J}{\partial u_{ij}} = 0, \Sigma_{i=1}^{c}u_{ij} = 1, j = 1, 2, \ldots, n,$

$$u_{ij} = \left[\frac{(x_j - v_i)^T\Sigma^{-1}(x_j - v_i)}{\Sigma_{s=1}^{c}(x_j - v_s)^T\Sigma^{-1}(x_j - v_s)}\right]^{-\frac{1}{m'-1}}, i = 1, 2, 3, \ldots, c \tag{20}$$

MFCM algorithm takes the following steps.

Algorithm 2: MFCM

Initialization: Given the number of categories c with $2 \le c \le p$, initialize the membership matrix with random function U, set the iteration counter t at 1, and define ε as the iteration stop threshold:

1: Calculate or update the cluster prototype with Eq. (19);

2: Calculate pseudo-inverse matrix of covariance through matrix inner product;

3: Calculate the value of an objective function with matrix inner product Eq. (17);

4: Set $t = t + 1$ to update the membership matrix U with Eq. (20);

5: If the value of the objective function calculated in Step 3 satisfies $J_{\text{MFCM}}t - J_{\text{MFCM}}t - 1 < \varepsilon$, stop the program and output the prototype matrix V and membership matrix U. Otherwise, proceed to Step 1.

The maximal value of the calculated membership degree decides the pattern to a certain cluster in MFCM algorithm. Mahalanobis distance is adopted to identify the relevant variables that significantly influence the clustering performance and reduce the error rate in the process of partitioning clusters with high attribute correlation.

4.2 Mahalanobis Distance-based Shadowed Rough-Fuzzy Clustering

Considering Mahalanobis distance, this paper proposes a shadowed rough-fuzzy clustering algorithm based on Mahalanobis distance for attribute related. The lower bound, the boundary region and Mahalanobis distance in a cluster are also advantageous and beneficial to produce better prototypes and more effective cluster partition. The objective function J_{MSRFCM} of MSRFCM is as same as that of FCM. In MSRFCM, prototypes are calculated by Eq. (6), and the calculation of membership degree is consistent with that of MFCM.

MSRFCM algorithm includes the following steps.

Algorithm 3: MSRFCM

Initialization: Given the number of categories c with $2 \leq c \leq p$, initialize the membership matrix with random function U, and set the iteration counter t at 1, and define ε as the iteration stop threshold:

1: Calculate the optimal β_i for each cluster based on shadowed set;

2: From β_i, determine the lower bound and boundary region for each cluster;

3: Calculate the prototypes with Eq. (6);

4: Calculate pseudo-inverse matrix of covariance through matrix inner product;

5: Calculate objective function value through matrix inner product with Eq. (17);

6: Set $t = t + 1$ to update membership matrix U with Eq. (20);

7: If the value of the objective function calculated in Step 5 satisfies $J_{\text{MSRFCM}}t - J_{\text{MSRFCM}}t - 1 < \varepsilon$, stop the algorithm. Then output the prototype matrix V, and the membership matrix U. Otherwise, proceed to Step 1.

5 Experiment

The experiments cover four intrusion data sets and intrusion detection clustering with FCM, SRFCM, MFCM, and MSRFCM. Matlab programming is used with initial parameters of all algorithms consistent with that of FCM algorithm. All intrusion detection experiments extract three subsets, each with a sample capacity of 2000 (including 1900 normal data and 100 abnormal attack data).

5.1 NSL-KDD Data Set

NSL-KDD is a modified version of the KDD CUP99, with some redundant traffic removed and imbalanced clusters structure improved. This data set was simulated using artificial data and generated in a closed network, where some of the networks involve proprietary network traffic with manual injected attacks. Among this data set, the training set includes 125,973 data records, and each record contains a class label attribute with a tag value of normal or attack, with nine discrete attributes and 32 continuous digital attributes, totaling 42. These types of attacks can be divided into four categories: Probing, DoS, U2R, and R2L.

The attributes in NSL-KDD data set include different data types. Direct experiment on raw data sets is inefficient and may influence the desirable outcome production. Therefore, data preprocessing is essential. All 42 attributes are selected for this experiment.

The three data subsets are selected from NSL-KDD data set, and the abnormal sample size of each set accounts for 5% of the total only. The overwhelming majority of normal data over intrusive data makes them valid for experiments. The sample structure is shown in Tab. 1 where different data sets have a respective abnormal type. Each sampled data set corresponds to respective attack type: DOS attack involving back, smurf, pod, and teardrop to data set 1; Probing attack involving ipsweep, nmap, portsweep, and satan to data set 2; and R2L attack involving ftp_write, guess_passwd, warezclient, and warezmaster to data set 3. As a small number of intrusion data related to the U2R attack type, such abnormal data with this attack type is not involved in this discussion.

The normalization process should apply to samples due to the great differences among the attributes of the records in the experimental data sets, so as to standardize the samples with different order of magnitude. Eq. (21) states the specific normalization process:

$$x'_{ij} = \frac{(x_{ij} - \overline{x_j})}{S_j}, i = 1, 2, \ldots, n. \tag{21}$$

where $\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, S_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2}$.

Table 1: Sample structure of NSL-KDD data set

Data set	Abnormal type	Number of abnormal samples	Number of normal samples	Total sample size
Data set 1	back, smurf, pod, teardrop	100	1900	2000
Data set 2	ipsweep, nmap, portsweep, satan	100	1900	2000
Data set 3	ftp_write, guess_passwd warezclient, warezmaster	100	1900	2000

Some indices, such as anomaly detection rate (Recall), Precision, and F1 score, are often used in evaluating the effectiveness of intrusion detection, as defined by Eqs. (22)–(24). Accuracy is also taken into consideration in Eq. (25).

$$\text{Recall} = \frac{\text{Number of abnormal records successfully detected}}{\text{Total number of abnormal data in the test data set}} \times 100\% \tag{22}$$

$$\text{Precision} = \frac{\text{Number of abnormal records successfully detected}}{\text{Total number of abnormal records calculated by algorithm}} \times 100\% \tag{23}$$

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{24}$$

$$\text{Accuracy} = \frac{\text{Number of records successfully detected}}{\text{Total number of data in the test data set}} \tag{25}$$

All the above indexes are used in the evaluation of network intrusion detection. Generally, the larger the value of Recall is, the more the intrusion data are detected. However, if more intrusion data are detected at the cost of more normal data misjudged as abnormal data, the corresponding detection performance may become worse. To thoroughly evaluate the intrusion detection performance, F1 score is put forward based on Precision and Recall. The higher the F1 score, the more effective intrusion detection.

Experiments are simulated by using the four algorithms, FCM, SRFCM, MFCM, and MSRFCM, in three subsets. The distribution of subsets and the clustering results are shown in Figs. 1–6, with their 23rd, 31st and 36th attributes where green and blue samples in the three data sets represent intrusion data and normal data, respectively.

In Figs. 1–6, comparing with the original distributions of data sets, the calculated prototypes through four different algorithms are close in value. In Figs. 2 and 4, MSRFCM partitions most normal data into the right cluster with the lowest partition error rate by comparing with the three algorithms of FCM, SRFCM, and MFCM. Thus, it achieves the highest detection performance. And compared with FCM, SRFCM and MFCM algorithms obtain the more real value-approaching prototypes in the abnormal cluster in Fig. 2 and partition more normal data into the right cluster with the lower partition error rate in Fig. 4, respectively. In Fig. 6, compared with the other three algorithms, MSRFCM obtains the most real value-approaching prototype in the abnormal cluster and can achieve the highest intrusion detection

value. In turn, SRFCM and MFCM algorithms obtain the more real value-approaching prototypes in the abnormal cluster than that of FCM.

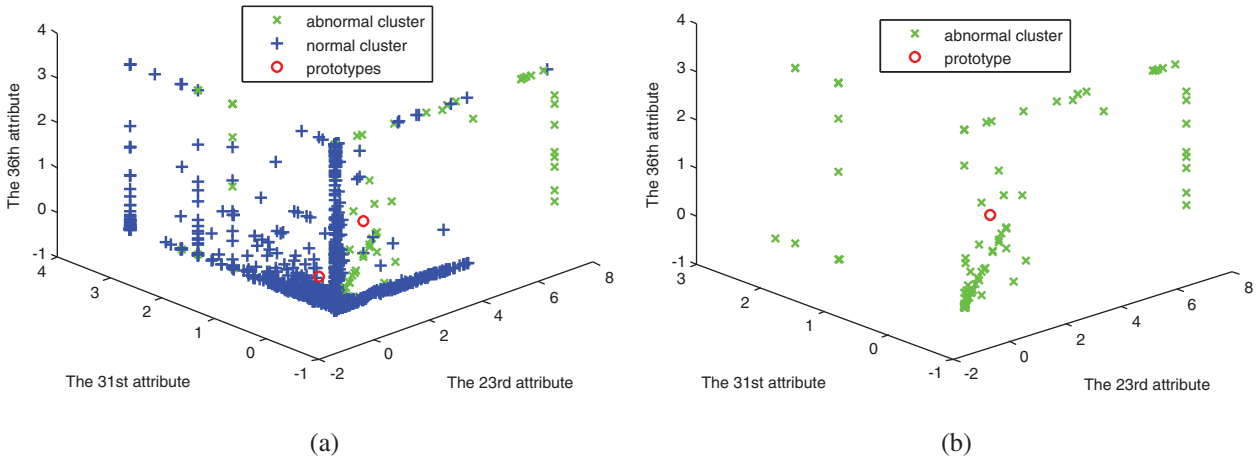

Figure 1: The distribution of data set 1. (a) Abnormal and normal clusters, (b) Abnormal cluster

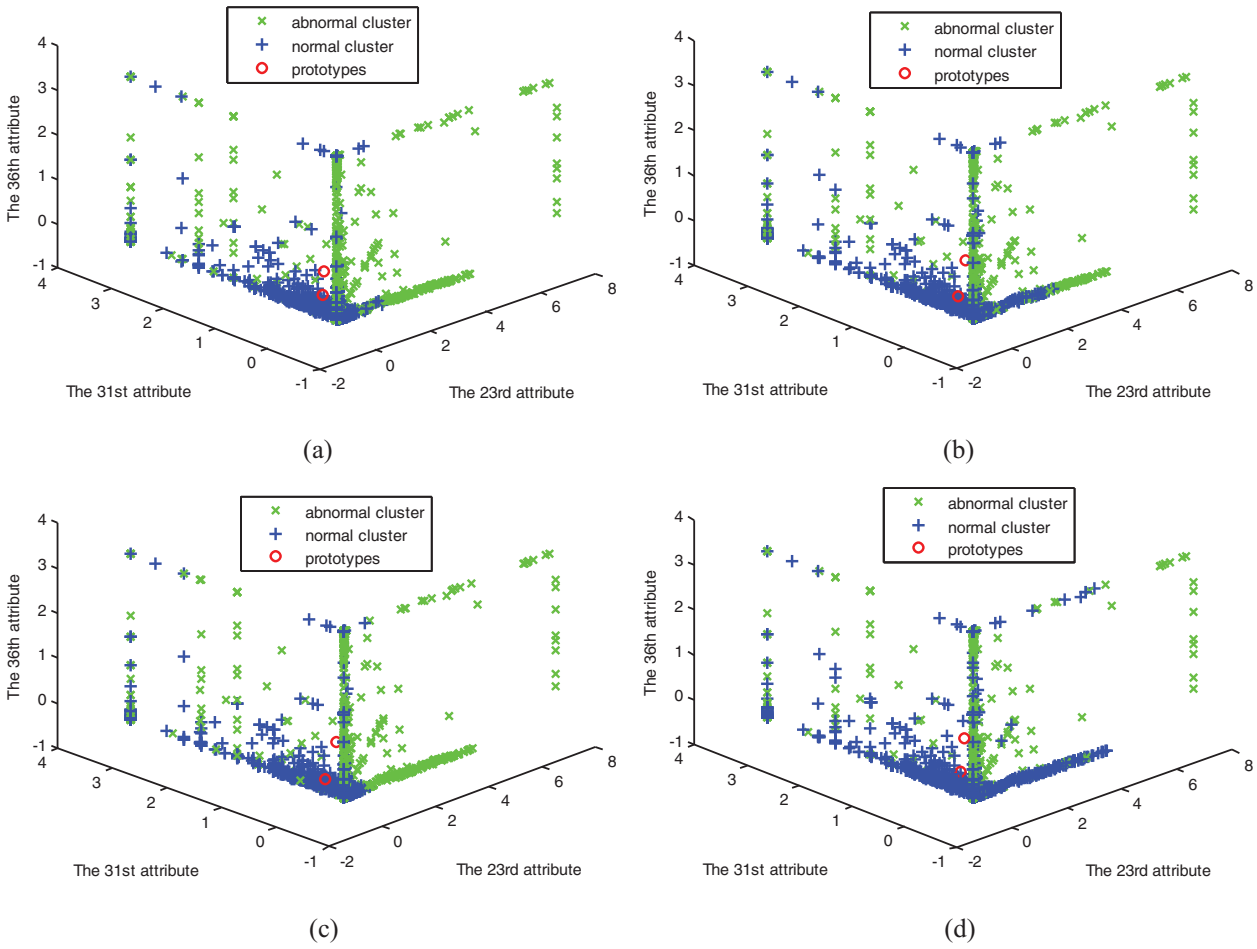

Figure 2: The clustering of four algorithms on data set 1. (a) FCM clustering, (b) MFCM clustering, (c) SRFCM clustering, and (d) MSRFCM clustering

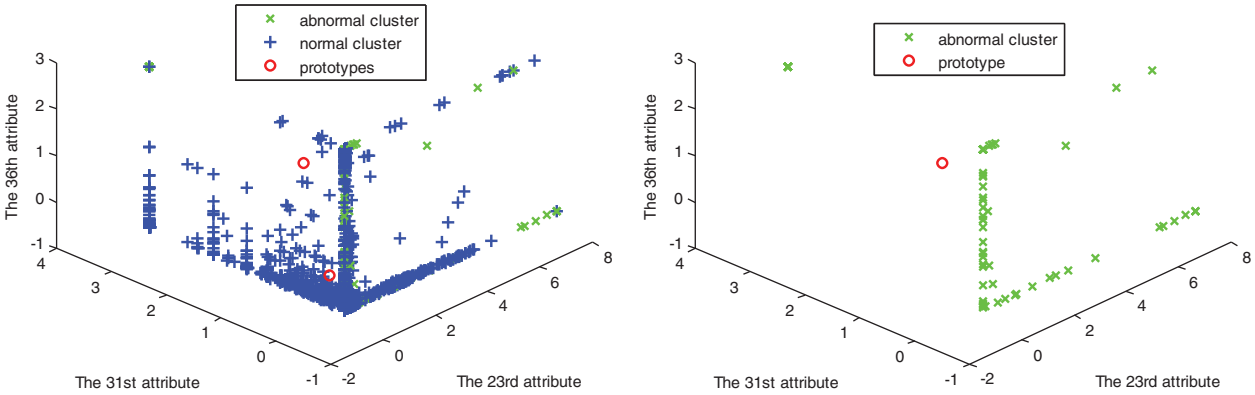

Figure 3: The distribution of data set 2. (a) Abnormal and normal clusters, (b) Abnormal cluster

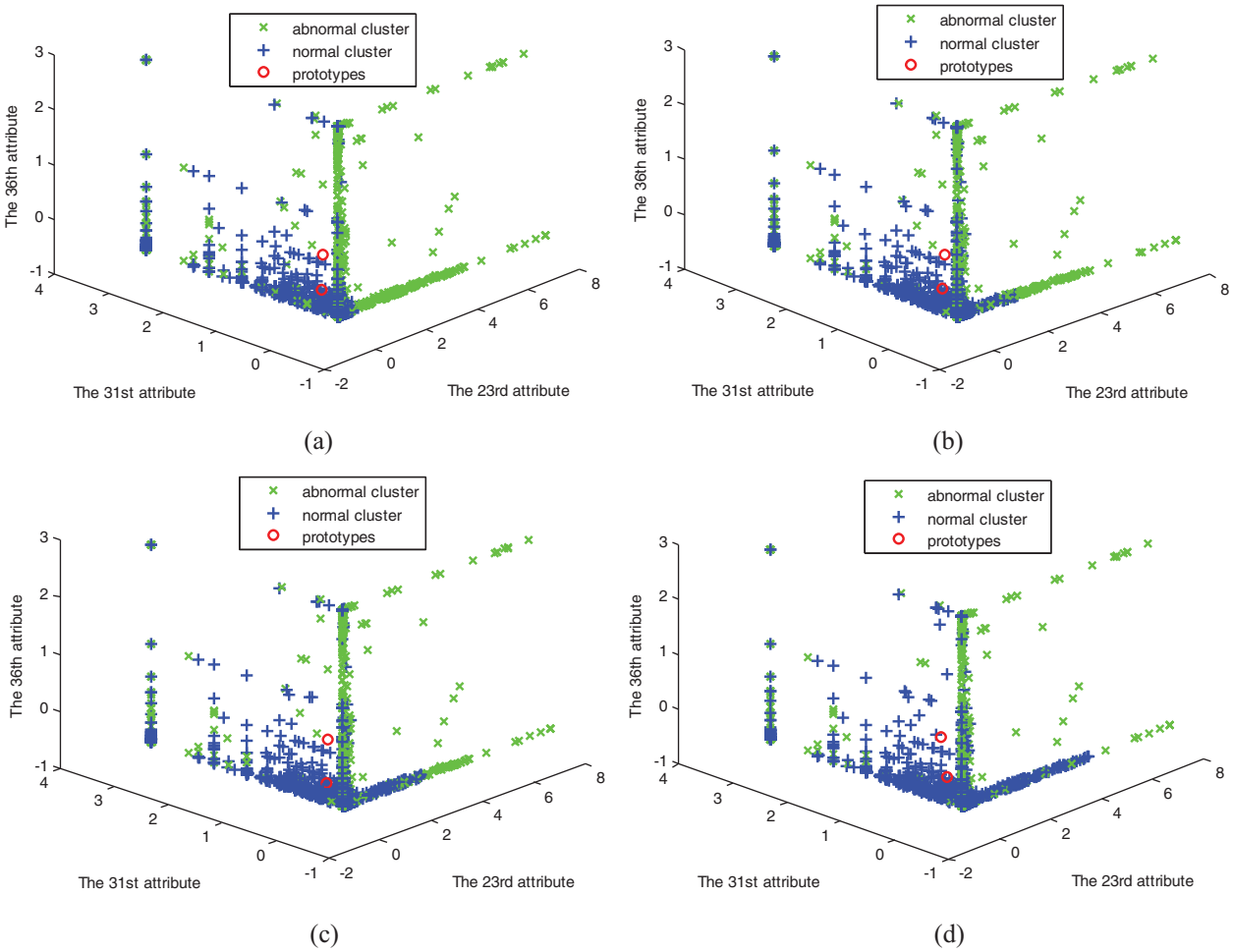

Figure 4: The clustering of four algorithms on data set 2. (a) FCM clustering, (b) MFCM clustering, (c) SRFCM clustering, and (d) MSRFCM clustering

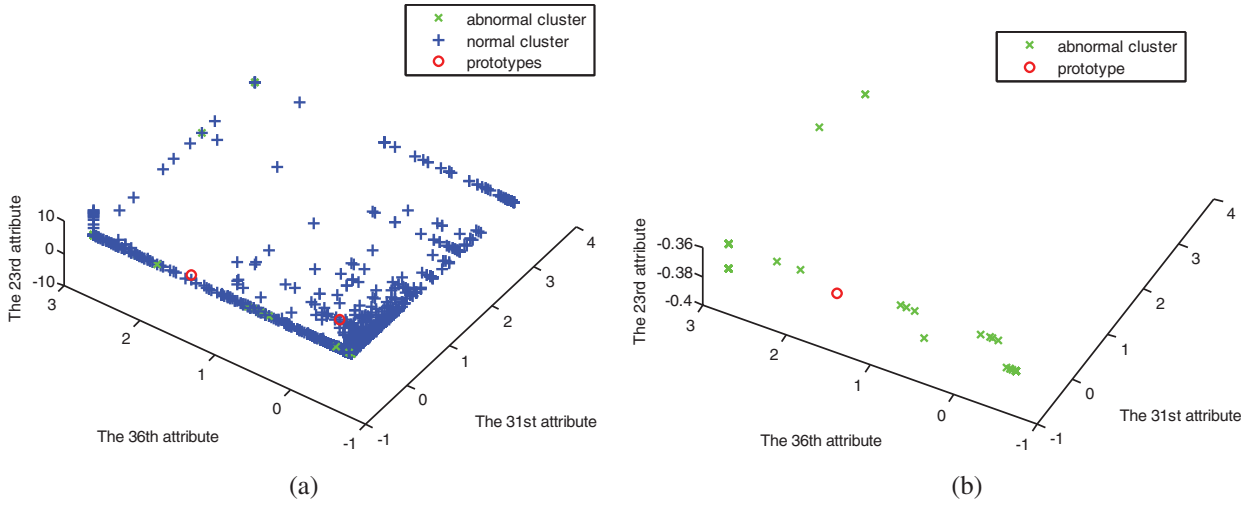

Figure 5: The distribution of data set 3. (a) Abnormal and normal clusters, (b) Abnormal cluster

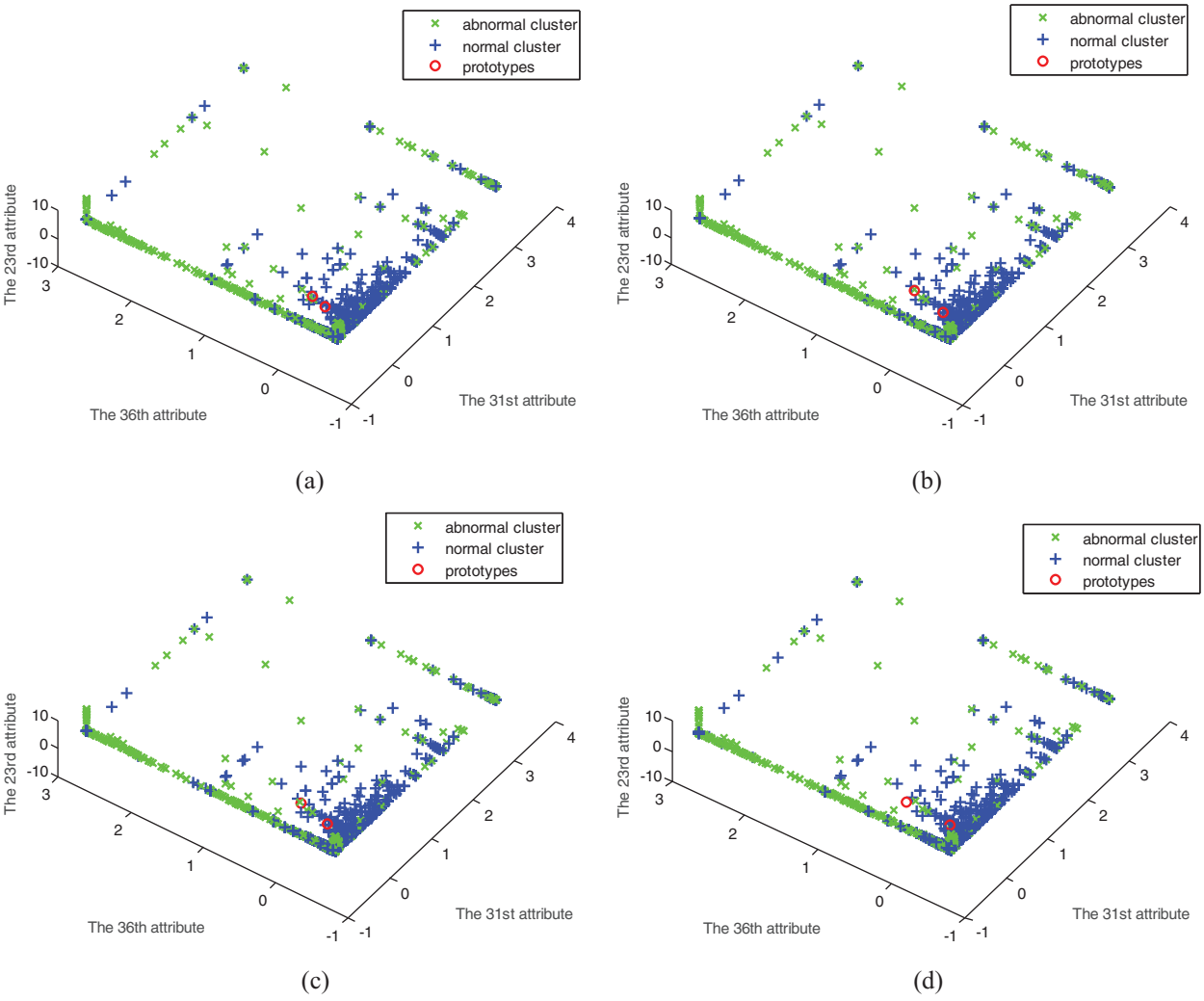

Figure 6: The clustering of four algorithms on data set 3. (a) FCM clustering, (b) MFCM clustering, (c) SRFCM clustering, and (d) MSRFCM clustering

Through statistical analysis of the experimental results in three data sets, the anomaly detection rate (Recall), Precision, Accuracy, F1 score, and the mean value of F1 score (MF1) of intrusion detection are calculated, as shown in Tab. 2.

Table 2: Comparison of detection performance in NSL-KDD data set

Algorithm	Data set	Recall (%)	Precision (%)	Accuracy (%)	F1	MF1
FCM	1	73	9.21	62.55	0.164	0.188
	2	91	10.76	61	0.192	
	3	100	11.59	61.85	0.208	
SRFCM	1	74	9.45	63.25	0.168	0.201
	2	88	12.68	69.1	0.222	
	3	99	11.93	63.4	0.213	
MFCM	1	73	10.72	68.25	0.187	0.204
	2	88	12.22	67.8	0.215	
	3	100	11.67	62.15	0.209	
MSRFCM	1	67	11.82	73.35	0.201	0.228
	2	87	15.88	76.3	0.269	
	3	96	12.08	64.85	0.215	

Based on Precision and Recall, F1 score is applied to evaluate the intrusion detection performance of different algorithms. In Tab. 2, in terms of F1 score, a series of Mahalanobis distance-based algorithms (MFCM and MSRFCM) are superior to FCM and SRFCM in the three data sets, respectively. In all data sets, MSRFCM clustering algorithm achieves the highest F1 value, and the mean intrusion detection F1 score (e.g., MF1) reaches the highest value of 0.228. The MSRFCM algorithm is superior to other algorithms on F1 score and MF1 score. As far as the accuracy is concerned, the experimental results show that FCM algorithm achieves the lowest accuracy value than the other three algorithms in three data sets. and MSRFCM algorithm acquires the highest accuracy value among all algorithms in three data sets. To sum up, MSRFCM algorithm is obviously superior to MFCM, SRFCM and FCM, and does yield very favorable outcomes in the intrusion detection experiments.

5.2 Other Data Sets

A 155-dimensional AWID data set [25] is provided with two versions: the one with labels corresponding to different attacks, and the other one with the attack labels organized into three major classes. Inside the attributes, some are useful for detecting attacks and others are just noise that may prove misleading. Meanwhile, three subsets are extracted with different attack types involved in each subset: the Flooding attack in data set 1, the Impersonation attack in data set 2, and the Injection attack in data set 3. The statistical analysis of the experimental results is shown in Tab. 3.

A 49-dimensional UNSW-NB 15 data set [25] is simulated in the Cyber Range Lab at the Australian Prototype for Cyber Security (ACCS). It is generated based on the combination of synthetic attack activity along with real modern normal behaviors, and the corresponding nine synthetic attack types are Backdoors, DoS, Analysis, Fuzzers, Generic, Worms, Exploits, Reconnaissance, and Shellcode, respectively. Three subsets sampled from it are corresponding to the respective attack type. The attack type set causing abnormal in data set 1 is comprised of Backdoors, Analysis, and Fuzzers. That causing

abnormal in data set 2 is comprised of DoS, Generic, and Exploits attack. The causing abnormal in data set 3 is comprised of Worms, Shellcode, and Reconnaissance. The statistical analysis of experimental results is listed in Tab. 4.

Table 3: Comparison of detection performance in AWID data set

Algorithm	Data set	Recall (%)	Precision (%)	Accuracy (%)	F1	MF1
FCM	1	0	0	83.1	0	0.140
	2	32	8.63	79.65	0.136	
	3	100	16.64	74.95	0.285	
SRFCM	1	100	17.12	75.8	0.292	0.306
	2	100	16.31	74.35	0.281	
	3	100	20.92	81.1	0.346	
MFCM	1	100	17.21	75.95	0.294	0.295
	2	98	16.93	75.85	0.289	
	3	100	17.73	76.8	0.301	
MSRFCM	1	100	17.24	76	0.294	0.321
	2	100	17.06	75.7	0.292	
	3	100	23.15	83.4	0.376	

Table 4: Comparison of detection performance in UNSW-NB 15 data set

Algorithm	Data set	Recall (%)	Precision (%)	Accuracy (%)	F1	MF1
FCM	1	99.67	33.71	70.55	0.504	0.501
	2	94.33	34.85	72.7	0.509	
	3	99.33	32.39	68.8	0.489	
SRFCM	1	94.33	45.21	82	0.611	0.545
	2	90.67	36.66	75.1	0.522	
	3	99.67	33.6	70.4	0.503	
MFCM	1	95.67	44.43	81.4	0.607	0.54
	2	94.33	35.78	73.75	0.519	
	3	99.33	32.86	69.45	0.494	
MSRFCM	1	99.67	46.43	82.7	0.634	0.558
	2	91.33	37.69	76.05	0.534	
	3	99.33	33.94	70.9	0.506	

An 85-dimensional CICIDS-2017 intrusion detection data set [25] is produced by the Institute of Network Security in Canada, with three subsets are sampled from it. Each subset corresponds to the respective attack type, Web attack to data set 1, Infiltration and PortScan to data set 2, and DDos attack to data set 3.The statistical analysis of experimental results is displayed in Tab. 5.

Table 5: Comparison of detection performance in CICIDS-2017 data set

Algorithm	Data set	Recall (%)	Precision (%)	Accuracy (%)	F1	MF1
FCM	1	11	1.80	65.60	0.031	0.095
	2	39	5.94	66.05	0.103	
	3	57	8.72	68	0.151	
SRFCM	1	9	2.56	78.30	0.040	0.127
	2	43	6.23	64.8	0.109	
	3	57	14.54	81.1	0.232	
MFCM	1	17	2.56	63.55	0.045	0.111
	2	43	6.2	64.6	0.108	
	3	57	10.71	74.1	0.180	
MSRFCM	1	17	2.72	65.45	0.047	0.130
	2	43	6.24	64.85	0.109	
	3	57	14.73	81.35	0.234	

The analysis on different algorithm performances in different intrusion data sets shows that MSRFCM has the highest F1 score and MF1 score, comparing with MFCM, SRFCM and FCM. The F1 value of MFCM algorithm in all data sets is higher than that of FCM algorithm. Comparing with the value derived from FCM algorithm, the deduced F1 value through SRFCM algorithm is higher in all data sets. For some instances, FCM algorithm has obtained F1 score of zero in data subset 1 of the AWID data set, and so it is unable to detect any intrusion behavior. By contrast, MSRFCM algorithm has an excellent performance in intrusion detection. Finally, the newly proposed algorithm has reflected its feasibility and advantages in intrusion detection through the experimental results.

6 Conclusion

This research explored the Mahalanobis distance-based SRFCM clustering algorithms to greater depth and elaborated on its applications in intrusion detection. The process and findings of the research are summarized as follows:

It analyzed and elucidated the effectiveness and feasibility of the improved algorithm in two steps. First, verify the effectiveness of the new method on NSL-KDD data set. Then, test its validity on preprocessed intrusion data selected from AWID, UNSW-NB 15, and CICIDS-2017 data sets.

In the discussion, SRFCM algorithm obtained real value-approaching prototypes based on iteration, and the dissimilarity measurement of Mahalanobis distance was used to identify the relevant variables and demonstrates its significant influence on the clustering performance and the error rate reduction in the process of partitioning clusters with high attribute correlation. With the merits of real value-approaching prototypes and the dissimilarity measurement of Mahalanobis distance, MSRFCM algorithm performed best among all of the algorithms. Besides, MSRFCM algorithm scored highest in intrusion data detection based on the simulation experiments on network intrusion data sets and the corresponding analyses of the Mahalanobis distance-based approach. In the future, coping with emerging security challenge on the Internet and combining various methods to integrate the advantages for detecting intrusion data is a worthy study.

Funding Statement: The authors would like to acknowledge the support of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai)(SML2020SP007) and The paper is supported by the National Natural Science Foundation of China (Nos. 61772280 and 62072249).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] K. Sha, W. Wei, T. Andrew Yang, Z. Wang and W. Shi, "On security challenges and open issues in Internet of Things," *Future Generation Computer Systems*, vol. 83, no. 6, pp. 326–337, 2018.

[2] I. Yaqoob, E. Ahmed, M. Habib ur Rehman, A. I. A. Ahmed and M. A. AI-garadi, "The rise of ransomware and emerging security challenges in the Internet of Things," *Computer Networks*, vol. 129, no. 12, pp. 444–458, 2017.

[3] U. Tariq, "Intrusion detection and anticipation system (IDAS) for IEEE 802.15.4 devices," *Intelligent Automation & Soft Computing*, vol. 25, no. 2, pp. 231–242, 2019.

[4] C. F. Cheang, Y. Wang, Z. Cai and G. Xu, "Multi-VMs intrusion detection for cloud security using Dempster-Shafer theory," *Computers Materials & Continua*, vol. 57, no. 2, pp. 297–306, 2018.

[5] J. Wang, J. Cao, R. S. Sherratt and J. H. Park, "An improved ant colony optimization-based approach with mobile sink for wireless sensor networks," *Journal of Supercomputing*, vol. 74, no. 12, pp. 6633–6645, 2018.

[6] T. Ling, L. Chong, X. Jingming and C. Jun, "Application of self-organizing feature map neural network based on k-means clustering in network intrusion detection," *Computers Materials & Continua*, vol. 61, no. 1, pp. 275–288, 2019.

[7] M. H. Luo, K. Wang, Z. P. Cai, A. F. Liu, Y. Y. Li *et al.,* "Using imbalanced triangle synthetic data for machine learning anomaly detection," *Computers Materials & Continua*, vol. 58, no. 1, pp. 15–26, 2019.

[8] J. Wang, Y. Cao, B. Li, H. Kim and S. Lee, "Particle swarm optimization based clustering algorithm with mobile sink for WSNs," *Future Generation Computer Systems*, vol. 76, no. 11, pp. 452–457, 2017.

[9] D. Li, H. Gu and L. Y. Zhang, "Fuzzy C-means algorithm with interval-supervised attribute weights," *Control and Decision*, vol. 25, no. 3, pp. 457–460, 2010.

[10] J. P. Mei and L. Chen, "Fuzzy clustering with weighted medoids for relational data," *Pattern Recognition*, vol. 43, no. 5, pp. 1964–1974, 2010.

[11] D. G. Yang, "Research of the Network intrusion detection based on fuzzy clustering," *Computer Science*, vol. 32, no. 1, pp. 86–87, 2005.

[12] S. Mitra, H. Banka and W. Pedrycz, "Rough-fuzzy collaborative clustering," *IEEE Transactions on Systems, Man, and Cybernetics (Part B)*, vol. 36, no. 4, pp. 795–805, 2006.

[13] P. Maji and S. K. Pal, "RFCM: A hybrid clustering algorithm using rough and fuzzy sets," *Fundamenta Informaticae*, vol. 80, no. 4, pp. 475–496, 2007.

[14] S. Mitra, "An evolutionary rough partitive clustering," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1439–1449, 2004.

[15] W. Pedrycz, "Shadowed sets: Representing and processing fuzzy sets," *IEEE Transactions on Systems, Man, and Cybernetics (Part B)*, vol. 28, no. 1, pp. 103–109, 1998.

[16] W. Pedrycz, "Granular computing—The emerging paradigm," *Journal of Uncertain Systems*, vol. 1, no. 1, pp. 38–61, 2007.

[17] J. Zhou, W. Pedrycz and D. Miao, "Shadowed set in the characterization of rough-fuzzy clustering," *Pattern Recognition*, vol. 44, no. 8, pp. 1738–1749, 2011.

[18] L. N. Wang and J. D. Wang, "Attribute weighting fuzzy clustering integrating rough set and shadowed set," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 4, pp. 1250010-1–1250010-25, 2012.

[19] J. H. Guo, D. Miao and J. Zhou, "Shadowed set based threshold selection in rough clustering," *Computer Science*, vol. 38, no. 10, pp. 209–210, 227, 2011.

[20] H. L. Wang, K. She and M. T. Zhou, "Shadowed set-based rough fuzzy possibility C-mean clustering," *Computer Science*, vol. 40, no. 1, pp. 191–194, 2013.

[21] S. M. Xiang, F. P. Nie and C. S. Zhang, "Learning a Mahalanobis distance metric for data clustering and classification," *Pattern Recognition*, vol. 41, no. 12, pp. 3600–3612, 2008.

[22] J. D. Peter and E. Peter, "A study of parameter values for a Mahalanobis distance fuzzy classifier," *Fuzzy Set and Systems*, vol. 137, no. 2, pp. 191–213, 2003.

[23] N. A. H. Haldar, F. A. Khan, A. Ali and H. Abbas, "Arrhythmia classification using Mahalanobis distance based improved Fuzzy C-Means clustering for mobile health monitoring systems," *Neurocomputing*, vol. 220, no. 12, pp. 221–235, 2017.

[24] X. M. Zhao, Y. Li and Q. H. Zhao, "Mahalanobis distance based on fuzzy clustering algorithm for image segmentation," *Digital Signal Processing*, vol. 43, no. 12, pp. 8–16, 2015.

[25] A. Aldweesh, A. Derhab and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," *Knowledge-Based Systems*, vol. 189, no. 1, pp. 1–19, 2020.

[26] J. Y. Cai, F. D. Xie and Y. Zhang, "New fuzzy clustering algorithms based on attribute weighted Mahalanobis distance," *Computer Engineering and Application*, vol. 48, no. 5, pp. 198–200, 2012.

[27] A. Ruiz and P. E. López-de-Teruel, "Nonlinear kernel-based statistical pattern analysis," *IEEE Transactions on Neural Networks*, vol. 12, no. 1, pp. 16–32, 2001.

[28] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 1–48, 2002.