Comparison between Support Vector Machine and Fuzzy C-Means as Classifier for Intrusion Detection System

Zuherman Rustam & Durrabida Zahras
Department of Mathematics, Universitas Indonesia, Depok, 16424, Indonesia

rustam@ui.ac.id

Abstract. In this globalization era, cybercrime has been entering every aspect through internet network. The development of Intrusion Detection System (IDS) is being studied deeply to solve the problem. There are several classifier algorithms for Intrusion Detection System such as Support Vector Machine (SVM) and Fuzzy C-Means (FCM). In this study, we will compare proposed model using both Support Vector Machine and Fuzzy C-Means to find a better result that increase accuracy of the network attacks. KDD Cup 1999 will be used to evaluate which algorithms work best. The results are very encouraging and show that SVM and FCM can be a useful tool for intrusion detection system. We found that SVM achieved 94.43% average accuracy rate while FCM achieved 95.09% average accuracy rate.

1. Introduction
When we compare our modern society with the early society, we can find that those are two different worlds. Because of the development of technology, our life has become more convenient and better. Internet has been the ultimate factor in driving globalization in recent years; many data can be outsourced entirely via Internet easily anytime and anywhere. The data can be anything; it can be password, thorough information regarding some system, people, companies, etc. Through the Internet, we can have immediate access to know everything that is happening in every corner around the world. With that convenience, it also attracts some parties to misuse the facility by putting some attack to our network. Intrusion Detection System is a method that can help us detect what kind of attack is trying to harm our data. It can monitor computer systems and network traffic. The increasing awareness of attack to the system triggered us to develop IDS with a better classifier.

In this paper, two classifiers are being studied. The first one is Support Vector Machine. It is a method that works in finding the best hyperplane that serves as a separator of two classes in the input space. The second one is Fuzzy C-Means. It is one of fuzzy clustering method that also works on classifying some classes using Euclidian space. We will use 10% “CORRECTED KDD CUP 1999 DATA” to see which classifier works best. This method known as powerful machine learning tools for classifications, application of SVM and FCM have been used by Rustam et al in[1,2,3,4] Intrusion Detection System itself, Support Vector Machine, Fuzzy C-Means, Experiment Results, and Conclusion will be discussed on the next section.

2. Intrusion Detection System
Intrusion Detection System is a method that monitors a computer system for activity that indicates access by unauthorized persons or computers. IDS will immediately report the attack once the attack is
detected. IDS is created to prevent attack entering the system. The attack detected by IDS can be divided into two categories: host-based attacks and network-based attacks. IDS itself can be divided in two ways based on the location in a network: the first one is Host based Intrusion Detection System (HIDS) and the other one is Network Intrusion Detection System (NIDS). HIDS can be classified into misused HIDS and anomaly-based IDS[5]. A misused HIDS detects intrusions by inspecting user unusual behavior such as the usage of the computer against the signatures of host-based intrusions. NIDS can be composed by a potentially large number of sensors, which monitor the traffic flowing in the network [6]. NIDS have encountered the challenge of detecting new attacks to the system. IDS play an important role in protecting our network.

There are several types of attacks using KDD Cup 1999 data that IDS has already classified into four categories. The four categories are:
1. Denial of Service (DOS). The type of attacks are Apache2, Back, Land, SYN Flood, Mail Bomb, Ping of Death, Smurf, Teardrop, etc.
2. Remote to Local (R2L), the type of attacks are Dictionary, Ftp Write, Guest, Imap, Named, Phf, Sendmail, Xlock, and Xsnoop.
3. User to Root (U2R), the type of attacks are Eject, Fbconfig, Fdformat, Loadmodule, Perl, Ps, and Xterm.
4. Probing Attacks (PROBE), the type of attacks are Ipsweep, Mscan, Nmap, Saint, and Satan.

3. Support Vector Machine
Support Vector machine is a machine learning algorithm used for classification and regression introduced by Vapnik in the late 1990s. Support Vector Machine or SVM is related to Structural Risk Minimization (SRM). At the first place, SVM is an initial form for binary classification, but now it can also be used for multiclass classification. SVM method does mapping form input space to a higher dimensional space to support nonlinear classification problems where a maximal separating hyperplane is constructed. Hyperplane is a linear pattern whose maximum margin gives the maximum separation between the decision classes.

3.1. SVM Characteristics[7]
Given dataset \( \{x_i, y_i\}_{i=1}^N \) where \( N \) is known as the number of samples, \( x_i \in \mathbb{R}^D \) is known as feature vectors from sample-\( i \), where \( D \) is the number of feature (dimension), and \( y_i \) is known as class labels. For two class classification problem \( y_i \in \{-1, +1\} \), however for the multiclass classification problem \( y_i \in \{1, 2, ..., k\} \) where \( k \) is the number of class. The main purpose of SVM is to find the best hyperplane:

\[
w \cdot x + b = 0
\]  

![Figure 1](image-url) SVM is trying to find the best hyperplane that separates the two classes, class A and B.
The optimization problem of SVM can be summarized as:

Minimize

\[
\frac{1}{2} ||w||^2
\]  

s.t. \( y_i (w^T \cdot x_i + b) \geq 1, \forall i = 1, ..., N \)  

(2)

The objective of (2) is to find \( w \in \mathbb{R}^n \) and \( b \in \mathbb{R} \) subject to (3), where \( w \) is set of weights and \( b \) is the bias. By solving the problem above, formula of \( w \) and \( b \) are obtained as follows:

\[
w = \sum_{i=1}^{N} a_i y_i x_i
\]

(4)

\[
b = \frac{1}{N} \sum_{i \in S} (y_i - \sum_{m \in S} a_m y_m x_m)
\]

(5)

and the decision function is:

\[
f(x) = w \cdot x + b
\]

(6)

which able to maximize the margins.

4. **Fuzzy C-Means**

Fuzzy C-Means is one of fuzzy clustering method introduced by Bezdek[8]. For a data set \( X = \{x_1, x_2, ..., x_n\} \subseteq \mathbb{R}^d \), we define \( n \times c \) Membership Matrix \( U = [u_{ij}], 1 \leq i \leq n, a \leq j \leq c \), and Cluster Center \( V = \{v_1, v_2, ..., v_c\} \) where each object in \( V \) is an element of \( d \)-dimensional Euclidean Space.

Mathematical model of Fuzzy C-Means method is given by:

\[
J(U, V) = \min \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^m d^2(x_i, v_j)
\]

(7)

With constraints:

\[
\sum_{j=1}^{c} u_{ij} = 1, i = 1, 2, ..., n
\]

(8)

\[
\sum_{j=1}^{c} u_{ij} > 0, i = 1, 2, ..., c
\]

(9)

\[
u_{ij} \in [0,1], j = 1, 2, 3, ..., c
\]

(10)

Where \( d \) is distance or dissimilarity function and \( m \in [1, \infty) \) is the fuzziness degree for cluster partition.

Cluster center and membership values are updated by using:

\[
v_j = \frac{\sum_{i=1}^{c} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m}, j = 1, 2, ..., c
\]

(11)

\[
u_{ij} = \left( \sum_{j=1}^{c} \left( \frac{d(x_i, v_j')}{d(x_i, v_j')} \right)^{\frac{2}{m-1}} \right)^{-\frac{1}{m-1}}, 1 \leq i \leq n
\]

(12)

In this research, we used Fuzzy C-Means (FCM) and the algorithm can be seen in Figure [2].

| Input: \( X, c, m, m_f, \epsilon, T \) | Output: \( U \) and \( V \) |
| --- | --- |
| 1. Initial condition: \( V^0 = \{v_1, v_2, ..., v_c\}, v_j \in \mathcal{C}_j \) | |
| 2. For \( t = 1 \) to \( T \) | |
| 3. \( m = m_t \) + \( \epsilon \) \( (m_f - m_t) \) | |
| 4. \( b = - \frac{1}{m-1} \) | Calculate membership |
| 5. \( u^1 = [u_{ij}], 1 \leq i \leq n, 1 \leq j \leq c \) by | |
using $u_{ij} = \frac{d^{2}(x_{i}, y_{j})}{\sum_{k=1}^{n} d^{2}(x_{i}, y_{k})}$, 1 ≤ $i ≤ n$, 1 ≤ $j ≤ c$

Update cluster:

6. $V^t = [v_1, v_2, ..., c]$ where $v_j = \frac{\sum_{i=1}^{n} u_{ij} x_i}{\sum_{i=1}^{n} u_{ij}}$, $j = 1, 2, ..., c$

7. if $E = \sum_{j} d^2(v_{jt}, v_{jt-1}) ≤ \varepsilon$
       STOP ELSE

8. Go to (2)

Figure 2. Fuzzy C-Means Algorithm

5. Experiments Results
We will make new classes using KDD99 “10% corrected data”[9]. Then we will use both Fuzzy C-Means and Support Vector Machine as classifiers. We divided the 10% data into 5 classes, the 5 classes are Denial of Service, User to Root, Remote to Local, Probing Attacks, and Normal.

Shown in Table 1 and Table 2 the results of classification into 5 classes using SVM and FCM

Table 1. Accuracy results of Support Vector Machines into 5 classes with RBF Kernel, parameter: $\sigma = 0.05$.

| No | Data Training (%) | Accuracy (%) | Running Time |
|----|------------------|--------------|--------------|
| 1  | 10               | 95.14        | 386.94       |
| 2  | 20               | 94.73        | 384.14       |
| 3  | 30               | 95.43        | 369.75       |
| 4  | 40               | 94.61        | 337.64       |
| 5  | 50               | 93.16        | 301.52       |
| 6  | 60               | 92.31        | 262.20       |
| 7  | 70               | 92.81        | 207.81       |
| 8  | 80               | 93.66        | 147.63       |
| 9  | 90               | 98.10        | 80.75        |
|    | Mean             | 94.43        | 275.37       |

Table 2. Accuracy results of Fuzzy C-Means into 5 classes with Euclidian distance.

| No | Data Training (%) | Accuracy (%) | Running Time |
|----|------------------|--------------|--------------|
| 1  | 10               | 95.35        | 112.86       |
| 2  | 20               | 95.68        | 109.41       |
| 3  | 30               | 96.33        | 103.05       |
| 4  | 40               | 95.67        | 97.69        |
| 5  | 50               | 94.55        | 85.98        |
| 6  | 60               | 92.39        | 73.75        |
| 7  | 70               | 93.34        | 58.14        |
| 8  | 80               | 94.77        | 42.08        |
| 9  | 90               | 97.78        | 22.58        |
|    | Mean             | 95.09        | 78.39        |
Shown in Table 3 and Table 4 the results of classification into 23 classes using SVM and FCM

**Table 3.** Accuracy results of Support Vector Machines into 23 classes with RBF kernel, parameter: $\sigma = 0.05$.

| No | Data Training (%) | Accuracy (%) | Running Time |
|----|-------------------|--------------|--------------|
| 1  | 10                | 90.33        | 1015.73      |
| 2  | 20                | 92.93        | 1017.31      |
| 3  | 30                | 93.28        | 969.27       |
| 4  | 40                | 93.92        | 904.95       |
| 5  | 50                | 91.88        | 783.30       |
| 6  | 60                | 93.03        | 671.80       |
| 7  | 70                | 95.93        | 541.20       |
| 8  | 80                | 96.84        | 383.55       |
| 9  | 90                | 98.79        | 211.20       |
|    | Mean              | 94.10        | 722.03       |

**Table 4.** Accuracy results of Fuzzy C-Means into 23 classes with euclidian distance

| No | Data Training (%) | Accuracy (%) | Running Time |
|----|-------------------|--------------|--------------|
| 1  | 10                | 89.28        | 8465.75      |
| 2  | 20                | 91.04        | 8452.34      |
| 3  | 30                | 90.91        | 8165.34      |
| 4  | 40                | 91.16        | 7618.78      |
| 5  | 50                | 89.85        | 6808.94      |
| 6  | 60                | 90.83        | 5904.78      |
| 7  | 70                | 94.89        | 4769.58      |
| 8  | 80                | 95.24        | 3398.38      |
| 9  | 90                | 98.49        | 1845.98      |
|    | Mean              | 92.41        | 6158.87      |

6. Conclusion
We use two different classifiers, which are Support Vector Machine and Fuzzy C-Means to compare the result. From our experiment, we can decide which classifier works best on each class. The use of Fuzzy C-Means on the tested data leads us to a better result on 5 classes with the average of its accuracy reach 95.09%. For 23 classes the winner goes to Support Vector Machines with the average of accuracy reach 94.10%.

References
[1] Z.Rustam and N. Olivera, Fuzzy Robust Kernel C-Means with Modified Kernel Nearest Neighbor Feature Selection for Intrusion Detection, The 3rd ISCPMS 2017, AIP Proceeding Conference. 2017
[2] Zuherman Rustam and Ni Putu Ayu Audia Arintari, Comparison between Support Vector Machine and Fuzzy Kernel C-Means as Classifiers for Intrusion Detection System using Chi-Square Feature Selection, The 3rd ISCPMS 2017, AIP Proceeding Conference. 2017
[3] Z. Rustam and A.S. Talita, Journal of Theoretical and Applied Information Technology, Fuzzy
Kernel C-Means Algorithm for Intrusion Detection Systems, JATIT&LLS, Vol. 81. 10th November, 2015.

[4] Rustam, Zuherman and Talita, Aini Suri. 2015. Journal of Theoretical and Applied Information Technology. Fuzzy K-Medoids Algorithm for Multiclass Multidimensional Data Classification, Vol 80, No 1. November 2015. Indonesia.

[5] Hu, Jiankun. 2010. Host-Based Anomaly Intrusion Detection. Australia Research Council Discovery Grant. Australia.

[6] Babatope, Longe Olumide; Lawal, Babatunde and Ayobami, Ibitola. 2014. Strategic Sensor Placement for Intrusion Detection in Network-Based IDS. MECS. Hong Kong.

[7] N. Cristianini and J.S. Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.

[8] N. B. and J. C. Bezdek, An Integrated Approach to Fuzzy Learning Vector Quantization and Fuzzy C-Means Clustering, IEEE Trans. Fuzzy Systems, Vol. 5, No. 4, 1997, pp. 622-628.

[9] KDD Cup 1999 Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html