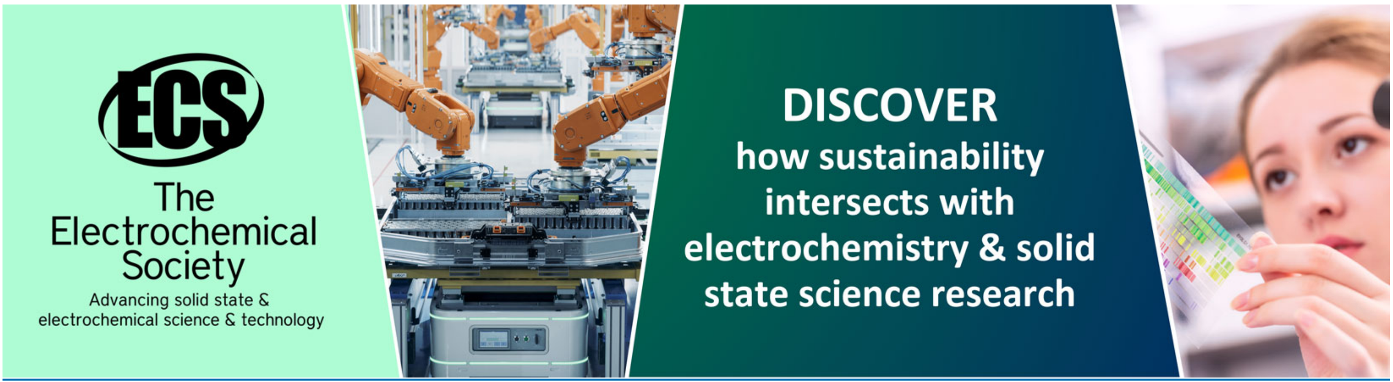
IOP Conference Series: Materials
Science and Engineering

PAPER • OPEN ACCESS

Comparison between fuzzy robust kernel c-means (FRKCM) and fuzzy entropy kernel c-means (FEKCM) classifier for intrusion detection system (IDS)

To cite this article: Nedya Shandri *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **546** 052071



Comparison between fuzzy robust kernel c-means (FRKCM) and fuzzy entropy kernel c-means (FEKCM) classifier for intrusion detection system (IDS)

Nedya Shandri[1], Zuherman Rustam[1*] and Devvi Sarwinda[1]

[1]Department of Mathematics, Universitas Indonesia Depok, 16424, Indonesia

*Corresponding author email: rustam@ui.ac.id

Abstract. Technology is growing very fast. We can now access everything using internet anywhere and anytime. That is why it is important to have internet security since we are always open to an online fraud, property damage and theft. IDS (Intrusion Detection System) can be used to detect any system or network attack. In this empirical study, we use dataset from KDD Cup 1999, which consist of five classes: normal, probe, dos, u2r and r2l. There is some classifier method for IDS, but in this study, we will use Fuzzy Robust Kernel C-Means (FRKCM) with Polynomial kernel and Fuzzy Entropy Kernel C-Means (FEKCM) with RBF kernel to find a better result that increase accuracy of the network attacks. There will be an accuracy comparison between FRKCM method and FEKCM method. The accuracy result from this study is 99% with time execution faster.

Keywords: Intrusion Detection System, Fuzzy Robust C-Means, Fuzzy Entropy C-Means, Kernel Function

1. Introduction

Technology is growing very fast. We can access everything using internet anywhere and anytime. The internet is group of small networks that connected to each other on a computer. Internet connection is very important since it allows us to access information and to communicate far easier. The Internet not only matters to businesses or citizens but also to government since it provides governments with an opportunity to function in a more innovative, engaging and cost-effective manner. However, this reliance on internet leads to an increasing number of cyber-attacks and data breaches, and numerous risks and challenges. One example of the threat is hackers [1]. Hackers can illegally gain access to a network and view the information on the local database, some of it highly confidential. The threat of hackers cannot be underestimated since now they are well structured, and the attacks might be undetected [2]. That is why it is important to have internet security to protect internal network. One of the tools we can use to prevent any system or network attacks is IDS (Intrusion Detection System).

IDS or intrusion detection is a system that can detect attacks from unauthorized users from other networks who want to try to get information from the network by checking the pattern of attacks on the computer network. However, IDS have many disadvantages as it cannot identify new attacks [3]. They most commonly detect known attacks based on defined rules or behaviour analysis through baselining the network. It can also cause system failure because when the IDS is turned off it will provide an opportunity for hackers to attack the system [4].

Nowadays, researches attempting to apply machine learning methods for IDS as solution to detect anomaly threats. Machine learning trains computer to process the information and act when required. Machine learning techniques enable computer to have thinking process like logical reasoning, trial and error and generalisations [5]. There are various machine learning algorithms that can be used for IDSs like Support Vector Machines, Decision Trees, Fuzzy Logic, Bayes Net and Naïve Bayes [2]. In this study, we use Fuzzy entropy kernel c-means (FEKCM) dan fuzzy kernel robust c-means (FRKCM) as classification methods. We will use 10% CORRECTED KDD CUP 1999 DATA to see which classifier works best.

Fuzzy C-Means is a clustering algorithm which solve classification problem through finding the most accurate cluster center. However, Fuzzy C-Means method can be interfered by the outliers since the membership must one [6]. The mechanism of the Fuzzy Robust C-Means and Fuzzy Entropy C-Means is the same with Fuzzy C-Means. In Fuzzy Robust C-means method, the outliers are force into a cluster [7]. In Fuzzy Entropy C-Means, an entropy measure works by identifying the total of the clusters and their center. This measure is different from other similar methods because after determining a cluster center, this measure does not revise values of all other data points.

The common problems in machine learning are the assumption that the data can be classified in linier. In fact, it is hard to separate data in linier, as stock data its self is a non linier data. Kernel function is needed as solution to this problem so the clustering process will run smooth and efficient. Kernel function is a function to represent the multiplication in a feature or high dimension room so the distance between data in one room can be calculated without transforming the data. In this research, algorithm fuzzy robust c means modified with kernel function, that is Fuzzy Kernel Robust C-Means.

2. Intrusion Detection System

IDS have been used to protect computer networks against both known and unknown attacks since 1970s [8,9,10]. IDS is a method that can detect attacks from unauthorized users from other networks who want to try to get information from the network by checking the pattern of attacks on the computer network. IDS itself can be divided into two ways based on the location in a network which are Host-based based Intrusion Detection System (HIDS) and Network-based Intrusion Detection System (NIDS) [11]. HIDS can be classified into misused HIDS and anomaly-based IDS [12]. A misused HIDS detects unusual activities of the computer that is suspected as intrusion based on prior information about specific attacks. NIDS consists of large number of sensors, which analyses data packets both inbound and outbound and offer real-time detection [13]. The challenge faced by NIDS is identifying new attacks to the system.

Based on KDD CUP 1999, the classes types of attacks as benchmark data for IDS research are classified into four categories [9]:

- Denial of Service (DOS) – type of attack that can shut down or weaken the power of the computer and makes the computer system crash and cannot operate well.
- Remote to Local (R2L) – the attackers send packages to find the weakness in the system and then act as local users to gain access.
- User to Root (U2R) – First attackers will access using normal account and then tries to find a weakness to get into root system to get super user privileges.
- Probing Attacks (PROBE) – The attackers scan the computer network to gain the information.

Table 1. Types of attacks of KDD CUP 1999

Classes Types of Attacks	Types of Attack
Denial of Service (DOS)	Apache2, Back, Land, SYN Flood, Mail Bomb, Ping of Death, Smurf, Teardrop
Remote to Local (R2L)	Dictionary, Ftp Write, Guest, Imap, Named, Phf, Sendmail, Xlock, and Xsnoop.
User to Root (U2R)	Eject, Ffbconfig, Fdformat, Loadmodule, Perl, Ps, and Xterm.
Probing Attacks (PROBE)	Ipsweep, Mscan, Nmap, Saint, and Satan

3. Methods

3.1 Fuzzy C-Means

Fuzzy C-Means (FCM) is expansion method from method K-means [13,14]. We can use Fuzzy C-Means (FCM) clustering techniques by assigning some membership values in the range of [0,1] to find a significant cluster [15]. The objective function of Fuzzy C-Means can be written as [13]:

$$\boldsymbol{J_m(U,V)} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{c}} (\boldsymbol{u_{ik}})^{\boldsymbol{m}} \parallel \boldsymbol{y_k} - \boldsymbol{v_i} \parallel_{\boldsymbol{A}}^{\boldsymbol{2}} \tag{1}$$

Where, c is number of cluster $(\boldsymbol{Y};\ \boldsymbol{2} \leq \boldsymbol{c} \leq \boldsymbol{n})$, $\boldsymbol{m}$ is weighting exponent $(\boldsymbol{1} \leq \boldsymbol{m} \leq \infty)$, U is fuzzy partition, v is vectors of center, $\boldsymbol{v_i}$ center of cluster $\boldsymbol{i}$.

The distance from y_k to v_i, calculated by:

$$d_{ik}^2 = \parallel y_k - v_i \parallel_A^2 = (y_k - v_1)^T A (y_k - v_1) \tag{2}$$

The function of fuzzy c-means with obstacles as follow:

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{c}} \boldsymbol{U_{ik}} = \boldsymbol{1}$$

Will be in optimal conditions if:

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m \boldsymbol{x}_k}{\sum_{k=1}^{n} (u_{ik})^m};\ 1 \leq i \leq c \tag{3}$$

$$U_{ik} = \left(\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)} \right)^{-1}\ ;\ 1 \leq \mathrm{k} \leq \mathrm{N}\ ;\ 1 \leq i \leq c \tag{4}$$

Vhere $\sum_{i=1}^{c} u_{ik} = 1 \forall k$, m is weighting exponent ($1 \leq m \leq \infty$), U is presented as membership matrix. $\boldsymbol{V} = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_c}\}$ are vectors of cluster centroids, $\boldsymbol{v_i}$ center of cluster $\boldsymbol{i}$.

3.2 Kernel Function

Suppose $\emptyset$ nonlinear mapping from input space $\mathbb{R}^d$ into feature space $F(\emptyset : \mathbb{R}^n \to for\ (x) \to \emptyset(x))$[15]. By using Kernel Fuzzy C-Means (KFCM) [15] and K is the kernel in the feature space [16]:

$$\boldsymbol{K(x,y)} = \langle \emptyset(\boldsymbol{x}), \emptyset(\boldsymbol{y}) \rangle$$

Where $\langle \emptyset(\boldsymbol{x}), \emptyset(\boldsymbol{y}) \rangle$ denotes the inner product operation. kernels allow computing inner products in the space, where one could otherwise not practically perform any computations because the inner function $\emptyset(\boldsymbol{x}), \emptyset(\boldsymbol{y})$ can be implicitly computed in F without knowing the mapping $\emptyset$ [15].

The distance between $\emptyset(\boldsymbol{x})$ and $\emptyset(\boldsymbol{y})$ define as [17]:

$$d(x,y) = \|\emptyset(x) - \emptyset(y)\| \tag{5}$$

$$d^2(x,y) = K(x,x) - 2(x,y) + K(y,y) \tag{6}$$

For $\boldsymbol{K(x,x)=1}$, so that $\boldsymbol{d^2(k,y)=2(1-K(x,y))}$.

With kernel function, Fuzzy C-Means [13] modified as:

$$J(\boldsymbol{U},\boldsymbol{V})=\sum\nolimits_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m(1-K(x_k,v_i)) \tag{7}$$

With Constraints:

$$\sum\nolimits_{j=1}^{c}u_{ik}=1,\ \ i=1,2,\dots,n \tag{8}$$

3.3 Fuzzy Entropy Kernel C-Means

This is a modification method of Fuzzy Entropy C-Means, which is sensitive to outlier and noise. To prevent this noise's effects, we will use kernel function that can decrease the outliers' effects and can be used for no separable data.

With kernel function, the modified function can be written as [18]:

$$J(\boldsymbol{U},\boldsymbol{V})=2\sum_{i=1}^{c}\sum_{k=1}^{n}t_{ik}(1-K(\boldsymbol{x_k},\boldsymbol{v_i}))+\frac{\sigma^2}{m^2c}\sum_{i=1}^{c}\sum_{k=1}^{n}(t_{ik}\log t_{ik}-t_{ik}) \tag{9}$$

Where $\boldsymbol{0\le t_{ik}\le 1}$, $\boldsymbol{c}$ is number of clustering, $\boldsymbol{n}$ is number of data points, t_{ik} is the typically of $x_k$in class i.

With,

$$\sigma^2=\frac{1}{n}\sum_{k=1}^{n}\sqrt{2(1-K(x,\bar{x}))} \tag{10}$$

is a normalization term with $\boldsymbol{\bar{x}=\frac{1}{n}\sum_{j=1}^{n}x_j}$.

Where:

$$t_{ik}=\exp\left(-\frac{2m^2c((1-K(\boldsymbol{x_k},\boldsymbol{v_i})+\lambda)}{\sigma^2+m^2c\lambda}\right),\forall i,k \tag{11}$$

$$v_i=\frac{\sum_{k=1}^{n}u_{ik}^m\boldsymbol{x_k}}{\sum_{k=1}^{n}u_{ik}^m} \tag{12}$$

For all i and $k\ge 1$ and x contains $c<n$ different data points.

3.4 Fuzzy Robust Kernel C-Means

Let $X=\{(x_i,y_i): i=1,2,\dots,m\}$ training data, where label y_i from $x_i\in R^n$, and dataset $X_j\subset X$ with $j=1,2,\dots,c$. There is $V=\{v_1,v_2,\dots,v_c\}$ and matrix $n\times c\ is\ u=[u_{ij}]$. Compliment function from membership u_{ij} $(f(u_{ij}))$ [8]:

$$f(u_{ij})=(1+u_{ij}lnu_{ij}-u_{ij}) \tag{13}$$

And objection function is defined as [8]:

$$J(\boldsymbol{U},\boldsymbol{V}) = \sum_{j=1}^{c}\sum_{i=1}^{n}[2u_{ij}^{m}(1 - K(\boldsymbol{x_i},\boldsymbol{v_j})) + \eta_i(1 + u_{ij}^{m}\ln u_{ij}^{m} - u_{ij}^{m})] \tag{14}$$

where $\boldsymbol{K}(\boldsymbol{x_i},\boldsymbol{x_i}) = \boldsymbol{K}(\boldsymbol{v_j},\boldsymbol{v_j}) = 1$.

will produce membership function defined as:

$$\boldsymbol{u_{ij}} = exp\left(-\frac{d^2(\boldsymbol{x_i},\boldsymbol{v_j})}{\eta_j}\right) \tag{15}$$

And prototype is updated by using:

$$\boldsymbol{v_j} = v_j + \alpha_t(x_i - v_j)exp\left(-\frac{d^2(\boldsymbol{x_i},\boldsymbol{v_j})}{\eta_j}\right) \tag{16}$$

With

$$\boldsymbol{\alpha_t} = \alpha_0\left(1 - \frac{t}{T}\right),$$

T is maximum iteration and t iterator.
The value of $\boldsymbol{\eta}$ calculated by:

$$\eta_j = mind^2(v_j, v_k), \boldsymbol{j} \neq \boldsymbol{k} \tag{17}$$

4. Datasets

In this study, the data used was Intrusion Detection System data from KDD CUP 1999. There were 494,021 samples, 42 features and labels containing information on 5 classes. In Table 2, features will be shown from the KDD CUP 1999.

Table 2. KDD CUP 1999 Data Features

No	Feature Name	No	Feature Name	No	Feature Name
1	duration	15	su_attemted	29	same_srv_rate
2	protocol_type	16	num_root	30	diff_srv_rate
3	service	17	nu_file_creations	31	srv diff host rate
4	flag	18	num_shells	32	dst_host_count
5	src_bytes	19	num_access_file	33	dst_host_srv_count
6	dst_bytes	20	num outbound cmds	34	dst_host same_srv_count
7	land	21	is_host_login	35	dst_host_diff_srv_rate
8	wrong_fragment	22	is_guest_login	36	dst_host_same_src_port_rate
9	urgent	23	count	37	dst_host_diff_host_rate
10	hot	24	srv_count	38	dst_host_serror_rate
11	num_failde_logins	25	serror_rate	39	dst_host_src_serror_rate
12	logged_in	26	srv_serror_rate	40	dst_host_rerror_rate
13	num_compromised	27	rerror_rate	41	dst_host_srv_rerror_rate
14	root_shell	28	srv_rerror_rate	42	attack_type

5. Results

In this section will show the results of accuracy, sensitivity and running time between FEKCM and FRKCM to solve IDS problems. We will ramdomize10000 data of each class. Formally, accuracy has the following definition [19]:

$$\text{Accuracy: } \frac{TP+TN}{TP+TN+FP+FN}$$

Where, TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives. Sensitivity can be expressed as:

$$\text{Sensitivity: } \frac{TP}{TP+FN}$$

Table 3, Table 4 and Table 5 show the accuracy and sensitivity achieved when FKEC and FRKCM with RBF kernel ($\sigma = 5\ and\ 1000$) was applied to 2 classes.

Table 3. Accuracy and Sensitivity FRKCM with RBF kernel ($\sigma = 5$)

		FEKCM		
Data	Data Training (%)	Accuracy (%)	Sensitivity (%)	Running Time (s)
Normal_DOS	90	100.00	100.00	0.33
Normal_U2R	90	93.33	90.91	4.72
Normal_R2L	90	100.00	100.00	18.84
Normal_PROBE	30	96.43	95.77	3.97

Table 4. Accuracy and Sensitivity FRKCM with RBF kernel ($\sigma = 5$)

		FRKCM		
Data	Data Training (%)	Accuracy (%)	Sensitivity (%)	Running Time (s)
Normal_DOS	90	100.00	100.00	0.23
Normal_U2R	90	100.00	100.00	0.14
Normal_R2L	70	93.33	96.43	0.55
Normal_PROBE	90	100.00	100.00	0.19

Table 3 and Table 4 show the accuracy achieved when FKECM and FRKCM $\sigma = 5$ was applied to the KDD CUP 1999. The highest accuracy with FEKCM (100%), sensitivity (100%) was obtained with 90% data training and a running time of 0.33 s. The highest accuracy with FRKCM (100%), sensitivity (100%) was obtained with 90% data training and a running time 0.19 s.

Table 5. Accuracy and Sensitivity FEKCM with RBF kernel ($\sigma = 1000$)

		FEKCM		
Data	Data Training (%)	Accuracy (%)	Sensitivity (%)	Running Time (s)
Normal_DOS	90	100.00	100.00	0.42
Normal_U2R	80	100.00	100.00	5.27
Normal_R2L	80	90.00	100.00	10.64
Normal_PROBE	50	97.00	97.96	7.88

Table 6. Accuracy and Sensitivity FRKCM with RBF kernel ($\sigma = 1000$)

		FRKCM		
Data	Data Training (%)	Accuracy (%)	Sensitivity (%)	Running Time (s)
Normal_DOS	90	100.00	100.00	0.19
Normal_U2R	90	100.00	100.00	0.14
Normal_R2L	70	93.33	96.43	0.50
Normal_PROBE	90	100.00	100.00	0.19

Table 5 and Table 6 shows the accuracy achieved when FKECM and FRKCM with $\sigma = 1000$ was applied to the KDD CUP 1999. The highest accuracy with FEKCM (100%), sensitivity (100%) was obtained with 90% data training and a running time of 0,42 s. The highest accuracy with FRKCM (100%), sensitivity (100%) was obtained with 90% data training and a running time 0.19 s.

Table 7 and Table 8 demonstrate accuracy for KDD CUP 1999 data using FKECM and FRKCM are slightly different (see table 7 and table 8).

Table 7. Accuracy FEKCM and FRKCM with RBF kernel ($\sigma = 5$)

	FEKCM			FRKCM		
Data	Data Training (%)	Accuracy (%)	Running Time (s)	Data Training (%)	Accuracy (%)	Running Time (s)
5 Class	20	73.68	20.16	90	91.11	0.95
23 Class	70	66.93	897.84	90	94.27	10.03

Table 8. Accuracy FEKCM and FRKCM with RBF kernel ($\sigma = 1000$)

	FEKCM			FRKCM		
Data	Data Training (%)	Accuracy (%)	Running Time (s)	Data Training (%)	Accuracy (%)	Running Time (s)
5 Class	20	74.52	30.63	90	91.11	0.97
23 Class	70	66.93	1102.48	90	94.27	9.56

From Table 7 and Table 8, we can see that FRKCM gives the best accuracy for 5 classes and 23 classes. The highest accuracy using FRKCM 94.27% with $\sigma = 1000$ resulted with 90% data training and a running time 9.56 s.

6. Discussion

In this research, we would like to compare FEKCM dan FRKCM method to solve ids problem. We will classified the data into two classes which are Normal_Dos, Normal_U2r, Normal_R2L, 5 classes and 23 classes. For each class we use n% (n= 10, 20,..., 90) data for training data and (n-100%) for testing data.

We will create table which consist of 5 classes: normal-DoS, normal-U2R, normal-R2L, and all. We will use Normal-DoS to determine whether the DoS is attacked or no. This is also applied to normal-Probe and the rest. While to determine a DoS attack, Probe attack, U2R attack, R2L attack or even not an attack at all we will use the class with label all.

From Tables 3, 4, 5 and 6, we can see that the accuracy and sensitivity from FEKCM and FRKCM in 2 classes achieved 100% for training data 90. But FRKCM gave better result since the time needed is only 0,19. While from 5 class and 3 class classification, there is a significant difference for accuracy which is 94% in 9.56 seconds.

7. Conclusion

The best classification of Intrusion Detection System data problem result is given from the FRKCM with RBF kernel. Thus, we have found the satisfying accuracy with rapid running time from this research. We could also continue using this research to find better method regarding IDS data classification problem which might obtain a better result.

Acknowledgements

This research was financially supported by University of Indonesia, with PIT.9 2019 research grant scheme (ID number NKB-0039/UN2.R3.1/HKP.05.00/2019)

References

[1] Dalziel, Henry and Willson, David, “Cyber Security Awareness for CEOs and Management,” Chapter 3, 2016. Pp 25-29.
[2] S. A. R. Shah and B. Issac, “Performance Comparison of Intrusion Detection Systems and Application of Machine Learning to Snort System,” Future Generation Computer Systems, 2018. Pp 157–170.
[3] W. Lee, et al., “ Real Time Data Mining based Intrusion Detection,” 2001.
[4] B. Setiawan, S. Djanali, and T. Ahmad, “A Study on Intrusion Detection Using Centroid-Based Classification,” Procedia Computer Science, 2017. Pp 672-681.
[5] Suricata, 2014. Available at: https://suricata-ids.org/ (Accessed: 23 March 2019).
[6] Krishnapuram, R., & Keller, J.M. (1993). A Possibilistic Approach to Clustering. IEEE Transactions on Fuzzy Systems. Vol.1, No.2. Pp 88-110.
[7] T.-N. Yang, S.-D. Wang, and S.-J. Yen, “Fuzzy Algorithm for Robust Clustering,” Proceeding of International Computer Symposium. Houilan. Taiwan, 2002.
[8] M. Ahmed, A. N. Mahmood and J. Hu, “A survey of network anomaly detection techniques,” Netw. Comput. Appl. 60, 2016. Pp 19-31.
[9] J. Hu, I. Khalil, S. Han, A. Mahmood, “Seamless integration of dependability and security concepts in soa: a feedback control system based framework and taxonomy,” J. Netw. Comput. Appl. 34 (4), 2011. Pp 1150–1159.
[10] W. Haider, et al., “Integer data zero-watermark assisted system calls abstraction and normalization for host based anomaly detection system,” 2015.

[11] Z. Rustam and Z. Durrabida, “Comparison between support vector machine and fuzzy c-means as classifier for intrusion detection system,” 2018.

[12] J. Hu, “Host-Based Anomaly Intrusion Detection,”Australia Research Council Discovery Grant. Australia, 2010.

[13] O. B. Longe, et al., “ Strategic Sensor Placement for Intrusion Detection in Network-Based IDS,” Hongkong, 2014.

[14] A.A. Rachman, and Z. Rustam, “Cancer Classification using Fuzzy C-Means with Feature Selection,” in Int. Conf. on Mathematics, Statistics, and Their Applications (ICMSA),pp. 31-34,2016.

[15] S. Subudhi, and S. Panigrahi, “Use of Optimised Fuzzy C-Means Clustering and Supervised Classifiers for Automobile Insurance Fraud Detection,” Computer and Information Sciences,2017.

[16] A. Wulan, M.V Jannati, Z. Rustam, and A.A Fauzan, “Application Kernel Modified Fuzzy C-Means for Gliomatosis Cerebri,” in Int. Conf. on Mathematics, Statistics, and Their Applications (ICMSA),pp.35-38,2016.

[17] A. Mekhmoukh, K. Mokrani, and M. Cheriet, “A modifed Kernelized Fuzzy C-Means algorithm for ois images segmentation: Application to MRI images,” International Journal of Computer Science Issues (IJCSI), vol. 9, No. 1, January 2012.

[18] F.H. Jun, W.X. Hung, M.H. Ping and W. Bin, “Fuzzy Entropy Clustering Using Possibilistic Approach,” Control Enginerring and Information Science, vol.15,pp.1993-1997,2011

[19] Rustam, Z., & Talita, A. S. (2015). Fuzzy kernel k-medoids algorithm for multiclass multidimensional data classification. *Journal of Theoretical and Applied Information Technology*, *80*(1), 147.