International Conference on Information Technology Systems and Innovation (ICITSI) 2014
Bandung-Bali, 24-27 November 2014
ISBN: 978-1-4799-6526-7

Fighting Cyber Crime in Email Spamming: An Evaluation of Fuzzy Clustering Approach to Classify Spam Messages

Arie Wahyu Wijayanto
School of Electrical Engineering and Informatics
Insitut Teknologi Bandung
Bandung, Indonesia
ariewahyu@students.itb.ac.id

Takdir
School of Electrical Engineering and Informatics
Insitut Teknologi Bandung
Bandung, Indonesia
takdir.rex@students.itb.ac.id

Abstract— **The rising of the modern Internet brought with it heap opportunities for attackers to gain illegal benefit from spreading spam mail. Spam is irrelevant or inappropriate messages sent on the Internet to a large number of recipients. Many researchers use a large number of classification method in machine learning to filter spam messages. But, there is still limited research which evaluate the use of clustering task in data mining to perform spam email segmentation. In this paper we endorse for fighting cyber crime by evaluating the fuzzy clustering approach in classifying spam emails using one of the most popular and efficient method in this field, Fuzzy C-Means. The experimental studies on public spam data set using various different parameter give promising result in this process.**

Keywords—spam classification; fuzzy clustering; email spamming; fuzzy c-means ; cyber crime;

I. INTRODUCTION

As the internet connection and society becoming our daily needs, the use of cyber technology in criminal activities is increasing significantly [1]. A major challenge is that the growing values of cyber crime data needs to be seriously and accurately investigated by all law-enforcement organizations [2]. Discovering and detecting cyber crime can similarly be difficult because busy network traffic and frequent online transactions produce large amounts of information, only a small portion of which identifies with illegal activities [2].

Email has turned into one of the quickest and most popular types of correspondence [3]. Email is additionally a standout among the most omnipresent and pervasive applications utilized consistently by a huge number of individuals around the world. Then again, the increase in email clients has brought about an significant increase in spam messages during the recent years. There are more than 3 billion email accounts over the world, and roughly 294 billion messages are sent for every day, about 78% of them are spam [4]. In other case, the process of denial of service attack may takes place when a person uses a large number of spam emails to attack an email account or server resulting in computer server crashes or delays in routing email [5].

In Indonesia, the use of spam email for crime is widely known. One of those cases occur in October 28, 2008. Figure 1 presented a sample header and content of case business mail spam which share a bulk information to end user. The attacker targeting innocent consumer by manipulating email address similar to some popular and bona fide company. Usually the attacker create a detailed and good quality content to convince consumers. The content may be consist of some deception which encourage to criminal activity.

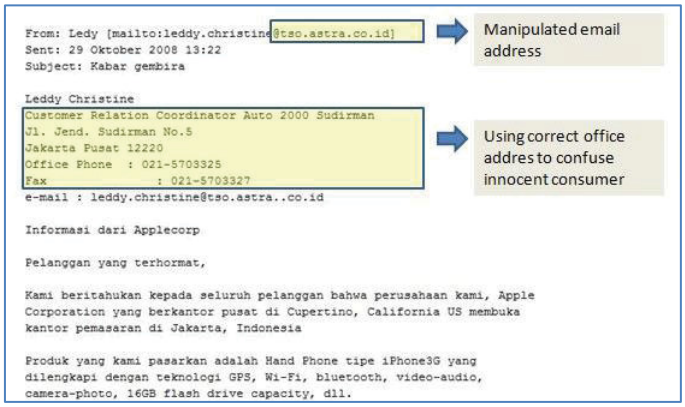

Figure 1. Sample of Indonesian business mail spam header and content [6]

In other hand, there are also many scam and other criminal activities spread by email spam. ESET analyzes first Android file-encrypting, TOR-enabled ransomware, are firstly distributed through spam emails [7], as presented in Figure 2. This spam email sent to random Android user email, which will be infects the phone. The malware also sends phone information, such as the IMEI number, to a server controlled by the attackers. The server itself uses encryption and sends communications through a number of difference servers to ensure it is extremely difficult to track users.

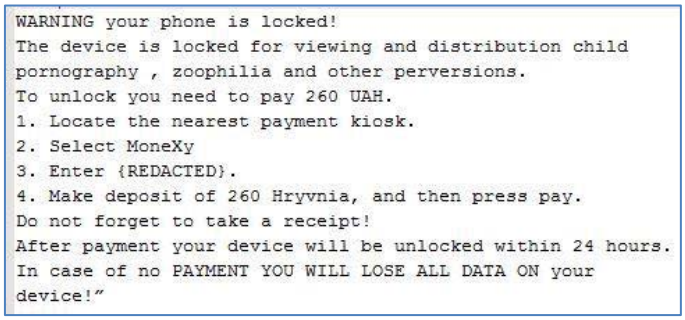

Figure 2. Other example of spam email case which contains malware related to Android infection through email [7]

II. Related Works

There are many researchs concerning about email spamming identification, classification and filtering activities. Rathi and Pareek [8] examine different data mining methods to spam dataset keeping in mind the end goal to figure out the best classifier for email classification. The authors analyze the performance of different classifiers with and without feature selection algorithm. Initially they explore different avenues regarding the whole dataset without selecting the peculiarities and apply classifiers one by one and check the results. At that point they apply Best-First peculiarity choice calculation to choose the desired features and afterward apply different classifiers for classification. It has been found that results are improved in terms of accuracy when they embed feature selection process in the experiment [8]. Finally they found Random Tree as best classifier for spam mail.

Xu and Yu proposes another spam filtering framework utilizing revised back propagation (RBP) neural network and automatic thesaurus construction [3]. The ordinary back propagation (BP) neural network has moderate learning velocity and is inclined to trap into a nearby least, so it will prompt poor execution and proficiency. The authors show the RBP neural network to defeat the confinements of the customary BP neural network. An well constructed thesaurus has been perceived as a profitable instrument in the powerful operation of content arrangement, it can likewise beat the issues in keyword-based spam filters that overlook the relationship between words.

It is a huge of research in email spam filtering and grouping using machine learning classification task. But there is still limited research in performaing the same goals using clustering task. In this research we focusing to fill those research gap by evaluating the clustering methods performance in same area, while the clustering methods is succesfully proven as effective method in other areas.

III. Theoretical Background

A. Cyber Crime

Numerous researchers concur that cyber crime is any illegal exercises performed through computer, however some differ on where it takes place [1]. There are many types of cyber crime such as computer hacking, internet fraud, unsolicited bulk mail (spam), credit card fraud, identity theft, online gambling, fraudulent websties, malware spreading, etc.

B. Email Spamming

Spam is unsolicited, unwanted, irrelevant or inappropriate messages sent electronically on the Internet to a large number of recipients [9]. These spam messages not just devour client time and vitality in distinguishing and uprooting the undesired messages, additionally cause numerous issues, for example, taking up the constrained post box space, squandering system data transmission and overwhelming paramount individual messages; besides, they can result in genuine harm to computers as machine infection [3]. The developing volume of spam messages and some of its related to cyber crime has brought about the need for more exact and effective spam separating framework.

C. Fuzzy Clustering

Fuzzy clustering is the procedure of assigning data items into a set of disjoint groups called clusters so that items in each cluster are more like one another to each other than items from diverse clusters [10]. The degree of membership in the fuzzy clusters relies on the closeness of the data item to the cluster centers [11]. Fuzzy c-means (FCM) algorithm is one of the most popular fuzzy clustering techniques because it is efficient, clear, and simple to actualize [10], [12].

IV. Fuzzy Clustering Approach

In this section we will present a fuzzy clustering approach to classify spam messages. Among other algorithms, the classical fuzzy clustering methods of Fuzzy C-Means (FCM) still widely used for classification task in many industrial areas. FCM was introduced by James Bezdek in 1981 [13], which is based on minimization of an objective function and is frequently used in pattern recognition [14]. To describe a method to determine the fuzzy c partition matrix μ for grouping a collection of n data sets into c classes, we define an objective function J for a fuzzy c -partition:

$$J = \sum_{k=1}^{n} \sum_{i=1}^{c} (\mu_{ik})^{m'} (d_{ik})^2 \quad (1)$$

Where μ_{ik} is the membership of the k-th data point in the i-th class. Runkler & Katz reformulate this formula into two different objective functions, respectively [15]:

$$J(V) = \sum_{i=1}^{c} \sum_{k=1}^{n} \frac{|v_i - x_k|^2}{\left(\sum_{j=1}^{c} \left(\frac{|v_i - x_k|}{|v_j - x_k|} \right)^{\frac{2}{m-1}} \right)^m} \rightarrow min \quad (2)$$

$$J(U) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \left| \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m} - x_k \right|^2 \rightarrow min \quad (3)$$

The objective of this proposed method is model accuracy, which can be measured by Partition Coefficient (PC), Classification Entropy (CE), Partition Index (SC), Separation Index (S), Xie and Beni's Index (XB), IFV index, Dunn Index (DI), and Alternative Dunn Index (ADI). Those measurement are usually used to measure the performance of clustering algorithms [16]. The PC index measures the amount of overlapping between clusters and for c number of cluster is defined as follows [16]:

$$PC = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{ij}^2 \quad (2)$$

where μ_{ij} is the membership of data point j in cluster i. The CE index measures the fuzziness of the cluster partition and is defined as follows [16]:

$$CE = -\frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{ij} \log (\mu_{ij}) \quad (3)$$

The optimal number of cluster is at maximum vaue of PC and minimum value of CE. The SC index is the ratio of the sum of compactness and separation of the clusters and is defined as follows [16]:

$$SC = \sum_{i=1}^{c} \frac{\sum_{j=1}^{N} (\mu_{ij})^m \|x_j - v_i\|^2}{N_i \sum_{k=1}^{c} \|v_k - v_i\|^2} \quad (4)$$

A better partition is indicated by a lower value of SC. The S index uses a minimum-distance separation for partition validity, while XB aims to quantify the ratio of the total variation within clusters and the separation of clusters.

$$S = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ij})^2 \|x_j - v_i\|^2}{N \, min_{i,k} \|v_k - v_i\|^2} \quad (5)$$

$$XB = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ij})^m \|x_j - v_i\|^2}{N \, min_{i,j} \|x_j - v_i\|^2} \quad (6)$$

The optimal number of clusters is where both of them are minimized [16], [17]. IFV are usually used as a validity function of fuzzy clustering for spatial data, because it's robustness and stability [18]. When IFV → max, the value of IFV is said to yield the most optimal of the dataset. It is defined as follows:

$$IFV = \frac{1}{C} \sum_{j=1}^{c} \left\{ \frac{1}{N} \sum_{k=1}^{N} u_{kj}^2 \left[log_2 C - \frac{1}{N} \sum_{k=1}^{N} log_2 u_{kj} \right]^2 \right\} \frac{SD_{max}}{\overline{\sigma_D}} \quad (7)$$

The maximal distance between centers is

$$SD_{max} = \max_{k \neq j} \|V_k - V_j\|^2 \quad (8)$$

The even deviation between each object and the cluster center is

$$\overline{\sigma_D} = \frac{1}{C} \sum_{j=1}^{c} \left(\frac{1}{N} \sum_{k=1}^{N} \|X_k - V_j\|^2 \right) \quad (9)$$

The Dunn Index (DI) is initially proposed to use at the identification of "reduced and decently differentiated groups". So the aftereffect of the bunching must be recalculated as it was a hard cluster algorithm [17]. The fundamental disadvantage of Dunn's index is computational since calculating gets to be computationally extremely sweeping as c and N increment. DI index is defined as follows:

$$DI = min_{i \in c} \left\{ min_{j \in c, i \neq j} \left\{ \frac{min_{x \in C_i, y \in C_j} d(x,y)}{max_{k \in c} \{ max_{x,y \in C, d(x,y)} \}} \right\} \right\} \quad (10)$$

The point of changing the original Dunn's index into the Alternative Dunn Index (ADI) was that the calculation gets to be more straightforward [17], when the dissimilarity function between two clusters ($min_{x \in C_i, y \in C_j} d(x,y)$) is evaluated in worth from underneath by the triangle-nonequality:

$$d(x,y) \geq |d(y,v_j) - d(x,v_j)| \quad (11)$$

where v_j is the cluster of the j-th cluster.

$$ADI = min_{i \in c} \left\{ min_{j \in c, i \neq j} \left\{ \frac{min_{x \in C_i, y \in C_j} |d(y,v_j) - d(x,v_j)|}{max_{k \in c} \{ max_{x,y \in C, d(x,y)} \}} \right\} \right\} \quad (12)$$

V. PROPOSED METHODS

In this section we proposed method in classifying spam email using fuzzy clustering approach as described in figure 3. It consists of three main phase: preprocessing, modelling, and evaluation phase. The modelling phase employ fuzzy clustering process FCM[11], where K-Medoid will be used as a benchmark due to its simplicity and popularity in clustering task. The evaluation phase consists of some validity index as described in the previous section and also the evaluation of accuracy.

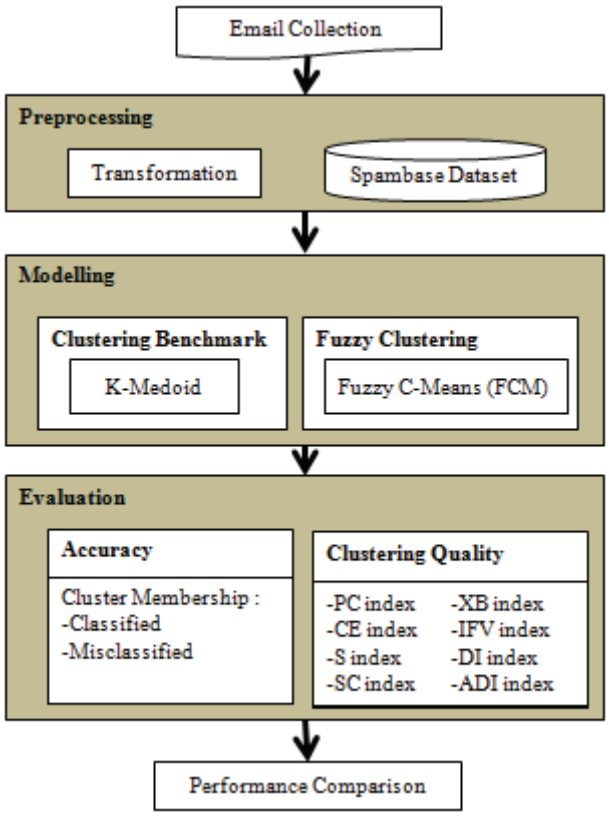

Figure 3. Workflow of Research Process

In preprocessing phase, the email collection were tranformed into clean data by performing these steps:

- Calculating the percentage of words and characters regarding spam characteristics as more detailed shown in Table 1 and its explanation.
- Calculating the number of capital letter and its sequence.

Those preprocessing phase has done in public data set from Spambase UCI Machine Learning data sets [19] which consist of 58 attributes and 4601 records. So, the reason to employ this data set is two fold. Firstly, the data set are available in public domain and has been widely used in many research on spam classification field, which provide the replicability aspect of this research. Secondly, the data set have preprocessed using qualified criterion regarding spam characteristics in order to simplify the clustering process on

this data. So, we can be more focused on the evaluation of clustering methods, than the feature selection process itself.

The last segment of Spambase data set signifies whether the email was considered spam (1) or not (0), i.e. spontaneous business email [19]. The greater part of the qualities show whether a specific word or character was habitually occurring in the email. The run-length characteristics (55-57) measure the length of groupings of sequential capital letters. Table 1 presents a detailed attributes of data set and its associatives data type and value.

TABLE I. ATTRIBUTES OF DATA SETS

Attributes Name	Data Type	Value	Attributes Type
1-48	Continuous Real	[0,100]	word_freq_WORD
49-54	Continuous Real	[0,100]	char_freq_CHAR
55	Continuous Integer	[1,...]	capital_run_length_average
56	Continuous Integer	[1,...]	capital_run_length_longest
57	Continuous Integer	[1,...]	capital_run_length_total
58	Nominal	[0,1]	class attribute

Here are the complete list of attributes type used in the simulation data set:

- type word_freq_WORD = percentage of words in the email that match WORD, i.e. 100 * (number of times the WORD appears in the email) / total number of words in email. A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.
- type char_freq_CHAR = percentage of characters in the email that match CHAR, i.e. 100 * (number of CHAR occurences) / total characters in email
- type capital_run_length_average = average length of uninterrupted sequences of capital letters
- type capital_run_length_longest = length of longest uninterrupted sequence of capital letters
- type capital_run_length_total = sum of length of uninterrupted sequences of capital letters = total number of capital letters in the email
- class attributes = denotes whether the email was considered spam (1) or not (0), i.e. unsolicited commercial email.

VI. EXPERIMENTAL RESULTS

In this section we evaluated the proposed method in classifying spam emails using experimental process. We have implemented the fuzzy clustering method in Matlab R2013a, executed under the environment of Intel Core i5-3210M CPU @2.50GHz, 4GB RAM and Windows 7 64bit operating system.

We have implemented the algorithm by improving the source code of Matlab Fuzzy Clustering and Data Analysis Toolbox [17]. We use public data set from Spambase UCI Machine Learning data sets [19] as described in the previous section. The Fuzzy C-Means are evaluated using various fuzzy exponent value againts kMedoid [17] algorithms as a benchmark. Some clustering parameters of those algoritms are set up as threshold $\varepsilon = 10^{-6}$, and number of clusters $c = 2$.

After performing a set of simulation using Spambase data set, we get the result of clustering task for 2 cluster. From the results, it is shown that Fuzzy C-Means using $m = 5.5$ give the best performance. Percentage of error and number of misclassified emails is minimum among the others, while kMedoid as a benchmark give worst which succesfully classified 2601 emails or 56.53 percent. The detailed results are presented in table 2.

TABLE II. COMPARISON OF CLUSTERING RESULT BY METHODS

Methods & Fuzziness Exponent	Number of Emails		Percentage (%)	
	Mis-classified	Classified	Mis-classified	Classified
FCM (m = 1,5)	1669	2928	36.27	63.64
FCM (m = 2,0)	1641	2960	35.67	64.33
FCM (m = 2,5)	1625	2976	35.32	64.68
FCM (m = 3,0)	1615	2986	35.10	64.90
FCM (m = 3,5)	1608	2993	34.95	65.05
FCM (m = 4,0)	1601	3000	34.80	65.20
FCM (m = 4,5)	1586	3015	34.47	65.53
FCM (m = 5,0)	1583	3018	34.41	65.59
FCM (m = 5,5)	1579	3022	34.32	65.68
FCM (m = 6,0)	1585	3016	34.45	65.55
FCM (m = 6,5)	1587	3014	34.49	65.51
FCM (m = 7,0)	1586	3015	34.47	65.53
k-Medoid	2000	2601	43.47	56.53

From the result above, we can conclude that for FCM is better at classifying spam emails, which it simulation using fuzziness exponent (m) equals to 5.5 shows the best result. Despite this result is not perfectly classify spam emails, the future of FCM implementation in this process is very promising due to its performance. The opportunity of enhancing other FCM parameters is possible to improve the classification accuracy.

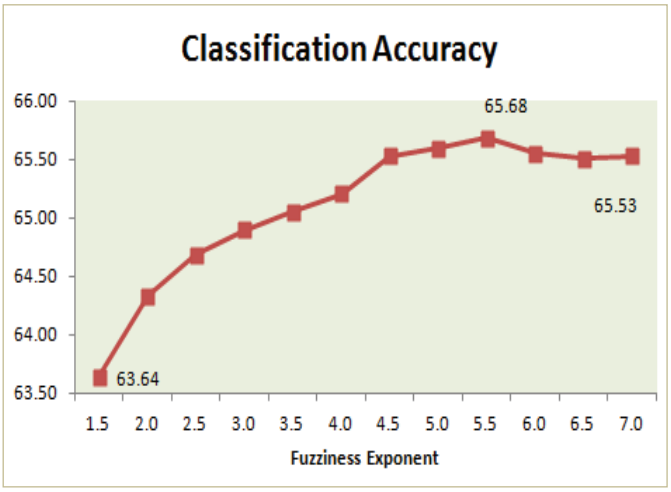

Figure 4. FCM Classification Accuracy Comparison

. In previous table, FCM is better than kMedoid. A detailed representative of FCM classification accuracy comparison is shown in Figure 6 Moreover as increasing the value of fuzziness exponent, the classification accuracy result against public data set is increasing too. When m = 5.5 the classification accuracy of FCM reach its best, and then decreased.

To evaluate the clustering quality of process resulted by FCM, in table 3 and 4 it is shown the validity index comparison. While performing simulation using spam base data set, it is shown that FCM with fuzzy exponent parameter equals to 5,5 reach a relatively better clustering validity based on its high value of IFV index and low value of XB index. It is shown a better clustering quality.

The result of PC index and CE index of m = 5.5 are not so good, but in [17] the authors stated that these two index has limitation in terms of decreasing and increasing due to the changing of cluster number. The SC index and S index value of m = 5.5 are lower, which also shown a better fuzzy clustering quality

TABLE III. COMPARISON OF FCM VALIDITY FOR M=1.5 TO M=4.0

Validity	Fuzziness Exponent *(m)* of FCM					
	1.5	**2.0**	**2.5**	**3.0**	**3.5**	**4.0**
PC	0.978	0.945	0.897	0.841	0.788	0.741
CE	0.038	0.097	0.185	0.276	0.354	0.418
SC	3.E-04	4.E-04	5.E-04	6.E-04	4.E-13	8.E-13
S	7.E-08	9.E-08	1.E-07	1.E-07	4.E-13	8.E-13
XB	44.467	46.999	34.751	30.589	74.049	31.547
IFV	2.E+04	8.E+03	4.E+03	3.E+03	1.E+08	5.E+07
DI	5.E-04	9.E-04	9.E-04	6.E-04	7.E-04	1.E-03
ADI	5.E-02	5.E-02	5.E-02	5.E-02	3.E-07	6.E-07

TABLE IV. COMPARISON OF FCM VALIDITY FOR M=4.5 TO M=7.0

Validity	Fuzziness Exponent *(m)* of FCM					
	4.5	**5.0**	**5.5**	**6.0**	**6.5**	**7.0**
PC	0.702	0.670	**0.644**	0.623	0.606	0.592
CE	0.468	0.507	**0.538**	0.562	0.581	0.597
SC	2.E-12	3.E-12	**4.E-12**	6.E-12	8.E-12	1.E-11
S	2.E-12	3.E-12	**4.E-12**	6.E-12	8.E-12	1.E-11
XB	17.118	10.771	**7.463**	5.536	4.328	3.524
IFV	3.E+07	2.E+07	**2.E+07**	1.E+07	9.E+06	8.E+06
DI	7.E-04	7.E-04	**9.E-04**	5.E-04	6.E-04	4.E-04
ADI	1.E-06	2.E-06	**2.E-06**	3.E-06	4.E-06	4.E-06

Despite there are no single perfect fuzziness exponent in this simulation, the use of FCM with m = 5.5 is the most reasonable and promising regarding its classification accuracy when handling spam emails and its clustering quality.

VII. DISCUSSION

Based on the research result, simulation, and related literature, there is a valuable remark to handle the cyber crime in spam email cases. According to current literature in preventing misleading of spam email, there are some guidelines to take after when utilizing email to publicize and inform others about any business or marketing related information:

- Not to give misleading header data
- Not to utilize deluding headlines
- Display clearly that the email is an advertisement
- Incorporate a legitimate area/postage address somewhere in the message
- Provide an opt-out option and address opt-out requests immediately

Due to experimental result, the enhancement of FCM fuzziness parameters is useful to increase the classification accuracy. There are some opportunity to more elaborate other clustering parameters in order to improve the accuracy. The use of metaheuristic optimization as described in the related work and theoretical background section from many research will be also possible to implement in FCM. All of those future work can be employed to gain a better performance of email spam classification using fuzzy clustering methods.

VIII. CONCLUSION

In this paper we encourage for fighting cyber crime by proposing the use of fuzzy clustering approach in classifying spam messages utilizing one of the most popular and effective method in this field, Fuzzy C-Means (FCM). We evaluated the fuzzy clustering method using various different clustering parameters against the classical clustering method as a benchmark, kMedoid. The experimental simulation using public spam base data sets give promising result in this process

for implementing fuzzy clustering approach in classifying spam email.

The result of this research can be more elaborated as a contribution in this field of study. Future work of this research may including the extension of FCM using metaheuristic optimization and other FCM parameter adjustment to enhance classification accuracy.

REFERENCES

[1] W. Chung, H. Chen, W. Chang, and S. Chou, "Fighting cybercrime: a review and the Taiwan experience," *Decis. Support Syst.*, vol. 41, no. 3, pp. 669–682, Mar. 2006.

[2] H. Chen, W. Chung, J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: A general framework and some examples," *Computer (Long. Beach. Calif).*, vol. 37, no. 4, pp. 50–56, 2004.

[3] H. Xu and B. Yu, "Automatic thesaurus construction for spam filtering using revised back propagation neural network," *Expert Syst. Appl.*, vol. 37, no. 1, pp. 18–23, Jan. 2010.

[4] Mashable, "The 41-Year History of Email," *Mashable.com*, 2012. [Online]. Available: http://mashable.com/2012/09/20/evolution-email/. [Accessed: 07-Oct-2014].

[5] CDPP, "Spam Attacks," *Australia's Federal Prosecution Service*, 2011. [Online]. Available: http://www.cdpp.gov.au/case-reports/spam-attacks/. [Accessed: 06-Oct-2014].

[6] BisnisDavit, "Contoh 3 Info Bisnis SPAM," 2008. [Online]. Available: http://www.bisnisdavit.com/info-bisnis/lain-lain/nopember-08/contoh--spam-3.htm. [Accessed: 07-Oct-2014].

[7] R. Lipovsky, "ESET Analyzes First Android File-Encrypting, TOR-enabled Ransomware," *WeLiveSecurity*, 2014. [Online]. Available: http://www.welivesecurity.com/2014/06/04/simplocker/. [Accessed: 07-Oct-2014].

[8] M. Rathi and V. Pareek, "Spam Mail Detection through Data Mining – A Comparative Performance Analysis," *Int. J. Mod. Educ. Comput. Sci.*, vol. 5, no. 12, pp. 31–39, Dec. 2013.

[9] S. J. Delany, M. Buckley, and D. Greene, "SMS Spam Filtering : Methods and Data," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9899–9908, 2012.

[10] H. Izakian and A. Abraham, "Fuzzy C-means and fuzzy swarm for fuzzy clustering problem," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1835–1838, Mar. 2011.

[11] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell USA: Kluwer Academic Publishers, 1981, p. 256.

[12] H. Izakian, A. Abraham, and V. Snasel, "Clustering categorical data using a swarm-based method," in *2009 World Congress on Nature & Biologically Inspired Computing (NaBIC)*, 2009, pp. 1720–1724.

[13] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy c-Means Algorithms for Very Large Data," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 6, pp. 1130–1146, Dec. 2012.

[14] M. Forouzanfar, N. Forghani, and M. Teshnehlab, "Parameter optimization of improved fuzzy c-means clustering algorithm for brain MR image segmentation," *Eng. Appl. Artif. Intell.*, vol. 23, no. 2, pp. 160–168, Mar. 2010.

[15] T. a. Runkler and C. Katz, "Fuzzy Clustering by Particle Swarm Optimization," in *2006 IEEE International Conference on Fuzzy Systems*, 2006, no. 3, pp. 601–608.

[16] G. Grekousis and H. Thomas, "Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The Fuzzy C-Means and Gustafson–Kessel methods," *Appl. Geogr.*, vol. 34, pp. 125–136, May 2012.

[17] B. Balasko, J. Abonyi, and B. Feil, "Fuzzy Clustering and Data Analysis Toolbox: For Use with Matlab," Veszprem, Hungary, 2005.

[18] C. Hu, L. Meng, and W. Shi, "Fuzzy clustering validity for spatial data," *Geo-spatial Inf. Sci.*, vol. 11, no. 3, pp. 191–196, Jan. 2008.

[19] K. Bache and M. Lichman, "UCI Machine Learning Repository." [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2013.