Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Profiling Internet Users' Activities using Fuzzy C-Means Algorithm

[1]Johnson Adeleke Adeyiga, [2]Kehinde Adebola Sotonwa, [3]Dosunmu Moyinoluwa

[123]Department of Computer Science and Information Technology, Bells University of Technology, Ota Ogun State, Nigeria.

[1]jaadeyiga@bellsuniversity.edu.ng, [2]kasotonwa@bellsuniversity.edu.ng, [3]amdosunmu@bellsuniversity.edu.ng

Article Info
Page Number: 7921-7933
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022

Abstract
The internet's openness has many benefits for consumers, like quick access, ease of use, and affordable software, but it also has drawbacks, such as being vulnerable to numerous cyberattacks that might harm businesses, governments, and individual individuals. Therefore, by looking at log files that display the pattern of users' actions in the network, profiling internet users using the Fuzzy C-Means (FCM) method is intended to give network managers a fast snapshot of the internet user's behavior within their managed network. Other attributes were applied to the data along with the FCM algorithm. MATLAB codes were used to implement the developed technique. The conventional FCM algorithm was used to assess the built system's performance in terms of the following metrics: sensitivity, precision, accuracy, specificity, and execution time. The investigation showed that the FCM algorithm performed better than the Simple K-Means (SKM) technique. Five (5) metrics were employed for the evaluation of the performance of the algorithms. The standard FCM showed slight improvement over the SKM. It was discovered that the standard FCM performed slightly better than the SKM with respect to all the metrics used in the evaluation apart from the execution time which can be trade off, since the system is intended to be used for control purpose and at such, accuracy is paramount. Hence, the high computational time of the standard FCM could be trade-off for better and accurate control system.

Keywords: -cyberattack; profiling; fuzzy; algorithm; network.

1. Introduction

The expansion of the internet brought about countless transactions and activities carried out electronically by individuals and businesses [6], and at this digital era, the use of internet produces more information than ever before. Understanding how users behave when they connect to the internet creates opportunities for better network design and security measures. The most important thing in the management of a network is to know the characteristics of the network user [11].

Illegal network usage cases range from pornography, internet fraud, trafficking illicit goods, intimidation and extortion, illegal intrusion to insult and slander, to mention but a few.

Therefore, it is imperative to profile user activities as it helps to discover hidden patterns in records to see clues that will assist in preventing cybercrime.

While there are many reliable, accurate and proactive tools used to analyze user activities, the most suitable for the cause of this research work is Log Analysis/Audit Logs. Log analysis is the key to threat detention [10]. It is the art and science of seeking computer generated records and making sense of it. Operating systems, apps, network hardware, and other intelligent or programmable devices all emit logs. With the help of a cluster algorithm, we can perform user profiling using the data obtained from this log analysis. The process of profiling can be supported by the clustering method. It is a method that enables one piece of data or data item to fit into a specific cluster according to predetermined criteria. By using this algorithm, the data/information provided can be divided into two or more groups or clusters.

Analyzing large data from log analysis can be very difficult because the daily activities of internet users will generate a very large amount of data. For this reason, a data mining technique would be appropriate in order to extract data, gain insightful information and establish a pattern around the data so as to give a meaningful insight to the data analyzed. Additionally, using a network's bandwidth for unrelated content like pornography or gambling consumes space; as a result, a system should be put in place to categorize and filter network usage to remove this unrelated data. Therefore, understanding internet user behavior is crucial for user profiling [6]. Therefore, the goal of this research is to use the Fuzzy C-Means (FCM) Algorithm to profile internet users.

2 Review of Related Works

There has been a good number of research works relating to the proliferation of web users and their activities on the internet. Some results have been made through the use of data mining techniques and clustering algorithms [6] did a survey analysis with a small sample of twenty students who have an academic background in social sciences. Questionnaires were used to gather different views and this were analysed to arrive at the behavioral profile of a cybercriminal. Although the research lacked the use of a cluster algorithm to proper classify the students, the result was able to provide a profile for the characterization cybercriminals through the combined analysis of the responses.

Reference [11] applied the K-means algorithm alongside log analysis to study the activities and behavior of internet users in an institution. Sample data was first extracted through log analysis. By separating the output, this log data included both the websites that people had accessed and the packets that were delivered and received across the network. The data was then preprocessed, and using programs that used the K-means technique, it was clustered into three categories depending on the frequency of website visits: low, medium, and high. The findings of this study were able to reveal details about the users based on their current online activity, but the study's data source for the profiling process had several drawbacks.

Reference [3] employed K-means clustering with minor modifications to accelerate the process of identifying patterns in crime and clustering algorithms was used to help detect

patterns in crime within a specific geography. To verify the findings, these strategies were used using actual crime statistics from a sheriff's office. In order to improve the forecast accuracy, semi-supervised learning techniques were also applied for knowledge discovery from the criminal records. The detectives' and other law enforcement personnel' productivity was increased because to this data mining framework. It was also used for homeland security and counterterrorism.

Another study by [1] incorporated several data mining and soft computing methods, using the traffic content of websites related to terrorism as the auditing information to trace and discover behaviors related to terrorism.

To extract the qualities and relations from web sites and recreate the scenario for crime mining, [7] had created a scenario in the meantime. To predict crime patterns, they employed a clustering/classification-based model, and to evaluate web data, they used data mining techniques. After implementing the K-Means clustering algorithm, the data was categorized into crime, none, and authentic users. The proposed work's accuracy was determined to be 94.75%, and it effectively detected bogus rate anomalies.

An investigation should be conducted after a crime is committed using a computer device to determine what happened based on some evidence. The primary goal of digital forensics is to gather precise evidence that establishes a defendant's guilt or innocence and to reduce erratic judgments brought on by the laborious nature of the inquiry. Due to this, [4] presented a fuzzy logic-based paradigm for digital criminal investigation. Due to the fact that all user transactions, activities, and hardware and software configurations are kept in one place, they were able to extract the essential evidence from Windows operating systems.

An effective two-stage cybercrime detection system was suggested in reference [8]; the system is adaptable to user and data behavior changes. The technique lowered the number of false alarms in addition to detecting cybercrime. Real-world data were generated on a small number of users, as well as synthetic data of many users with a variety of usage patterns. Experiments were then run on the two sets of uniquely generated data. The K-means clustering technique was used to divide the collected data into three categories. The remaining datasets were utilized for prediction and cybercrime detection, and 75% of the real data and fraud history data were used for training. The best feature that distinguished the provided samples was taken from the training set using Bootstrap ID3, a decision tree technique. The proposed work's accuracy was 94.67%, and it effectively finds bogus rate anomalies.

Based on the notion that an intruder's conduct can be divided into various stages that are active at various periods. The use of short-term fuzzy profiles and contexts from individual intrusion phases by [1] led to improved low-level attack detection accuracy. A context-driven, flexible implementation framework built on a double layer hierarchy of fuzzy sensors is the end product.

[9] put forth a two-phase clustering approach that used large datasets to automatically identify comparable case subgroups. The computation of numerous qualities that connect to a

criminal offender's behavior trait in the categorization domain employed the information gain ratio (IGR) ideas. The clustering procedure was utilized to locate related case subsets using the findings from this weighing phase. Additionally, [13] binary clustering and classification techniques were used to analyze criminal data with the goal of identifying criminals based on witnesses or other hints found at the crime scene. An auto correlation model was additionally employed to validate the offender. The work's weakness is that it does not take into account a scenario in which there are no witnesses or hints at the crime site.

It's also crucial to remember that the majority of internet network attacks are typically recorded in log files with a specified data structure. Based on this, [5] made an effort to streamline the process of discovering and recognizing illicit online attacks. The identifying procedure was then aided by the use of the clustering technique. After utilizing the K-means clustering technique to group the data log file into three categories of attacks, the data was employed in the network.

[2] K-Mean clustering algorithm and data mining techniques were used to analyze criminal online data resulting from social media. According to the results, crime data mining holds great promise for improving the efficacy and efficiency of criminal and intelligence investigation.

3. Research Approach

This section outlines the framework of the research methodologies used to achieve the goals and objectives of the study. The Fuzzy C-Means (FCM) algorithm was employed in this study to profile internet users. The Simple K-Means (SKM) method was also used in this process. To allow for a full and in-depth review of the outcomes, two alternative clustering algorithm variants—a hard clustering algorithm and a fuzzy clustering algorithm—were chosen. This was accomplished by carrying out the procedures in the block diagram in Fig. 1. Data collection is the first step, which is then utilized to evaluate the system's efficacy. Data cleaning algorithms were used to fill in missing values, spot outliers, and fix inconsistencies before the data was further processed. The raw data was then replaced with higher level concepts using concept hierarchies, transforming the data into the right form needed to apply the generalization concept. Feature Scaling Techniques were used to extract and normalize the attributes. Then, other methods credited with knowledge generation were implemented alongside the FCM procedures. The profiling system was put into use in MALAB 2015 to create various clusters. Accuracy, Specificity, Precision, and Execution Time were used to assess the results produced by the various clusters that were constructed.

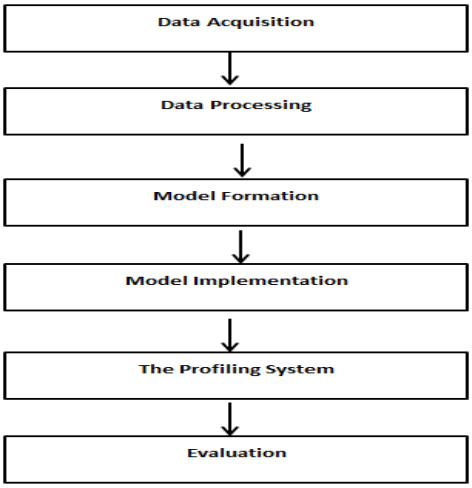

Fig. 1: Block diagram of the profiling system

3.1 Description of the model

The required log files are obtained from www.kaggle.com/hassan06nslkdd, a standardized data set, as part of the data collection process. There was created a data warehouse. The profile that was developed was subjected to the FCM algorithm in order to generate various clusters, and the knowledge deduced from the various clusters formed will protect organizations. Based on their behavioral inclinations, the clusters that were created were able to group the users. The suggested mode is depicted in Fig. 2.

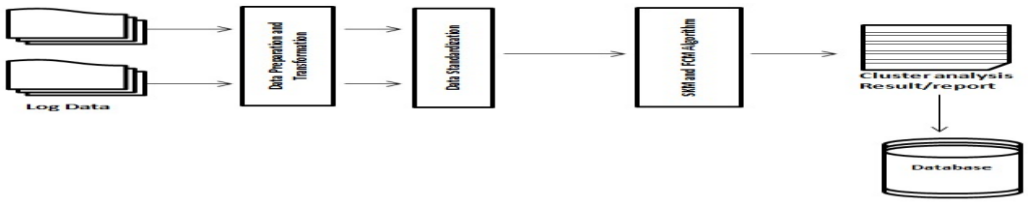

Fig. 2: The Model

3.2 Data acquisition, preprocessing and standardization

The data used consists of recorded events that occur in an operating system, or server/network. The data set is information about a device's activity or documents from web pages on the internet that related to user/device, was downloaded online which involves acquiring relevant log files ad is available at www.kaggle.com/hassan06nslkdd. And it was used to test the performance of the FCM clustering algorithms for the users' behavioral activities.

Using these data, we may utilize clustering techniques to profile user activity into existing clusters. Prior to doing standardization to extract the required data for mining, preprocessing was done on the collected data to clean, integrate, and transform the data. Before choosing the necessary attributes needed for profiling, some inconsistencies in the data were corrected by filling in some missing variables. To do this, the exported data was cleaned (unnecessary data was removed) and converted to the CSV (comma separated values) format using Microsoft Excel. Data cleaning routines were used to fill in missing values, identify outliers,

and correct data inconsistencies. The raw data was then replaced with higher level concepts utilizing concept hierarchies, transforming the input into the appropriated form required for concept generalization. By employing the vector slice approach to remove as much unnecessary and redundant information as feasible, it was further reduced through feature selection.

Since FCM clustering algorithm worked with integers and floats and the data set consists of IP addressed, it would be important to standardize the data to ensure centroids are properly calculated and to produce accurate results.

3.3 Attribute selection/normalization

The feature scaling concept was used to normalized the data after preprocessing, where all the various attributes have been converted into numeric value and the range of the feature of the data were reduced to a scalar. The relevant attributes needed to profile the internet users, such as, devices IP address, source IP address (website), time spent on the website, and number of times visited, will then be extracted from the standardized data in dot csv by the clustering program. This is required because different variables in the acquired data set have varying ranges, and it is expected that all data would fall within the same range. Equation 1 was used to determine Z, the normalized value of one of the observed values of x, and this was accomplished.

$$Z = \frac{x=\min(x)}{\max(x)-\min(x)} \qquad (1)$$

where x's lowest and maximum values are determined by its range.

3.4 Criteria for standardizing data items

Focused mainly on the source IP address, destination IP address, time spent and the link of this destination addresses e.g. google.com. for the data items of these attributed to be understood by the clustering algorithm, it had to be converted into plan integer values. Shown in Table 1 is the criteria for standardizing data items used. While each IP address that represented a user on the network was assigned a number sequentially, the time was simply converted from milliseconds to seconds.

Table 1: Criteria for standardizing destination address

Website Type	Value Assigned
Educational Search Engines News System Related Websites Software Update	1-3

Music Download Social Media Betting	4-6
Torrent Pornography Copyright infringing	7-10

1.4.1 The SKM algorithm

With SKM clustering, n objects are divided into k clusters, and each object is assigned to the cluster with the closest mean. The maximum number of distinct clusters produced by this method is k. Since it is unknown as a prior, the optimal number of clusters k that will result in the maximum separation (distance) must be calculated from the data. SKM clustering aims to reduce the squared error function, or total intra-cluster variance:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - C_j\|^2 \qquad (2)$$

Where J is the objective function, k is the number of clusters, n is the number of cases, x is the case i, ‖ is the distance function and C is centroid for cluster j.

The SKM clustering technique was used in this investigation and performed as expected. In the initial phases, main data collected includes details about the websites users have browsed online. Along with information on websites visited, it also includes information about changes to the operating system, web browser, and adverts that typically crop up. This is how the algorithm works:

Data is first clustered into k groups, where k is a predetermined number.

Step 2: Randomly choose k points to serve as cluster centers.

Step 3: Euclidean distance function is used to assign things to their closet cluster.

Step 4: Determine the centroid of each cluster's objects.

Step 5: Repeat steps 3 and 4 as necessary to give each cluster the same number of points across successive rounds.

1.4.2 The FCM clustering algorithm

Phase 1: Using fuzzy partition, this step computes the membership matrix (U).

Phase 2: The centroids, which serve as the cluster analysis system's nucleus, are computed in this step.

$$C_i = \frac{\sum_{j=1}^{n} U_{ij}^{m} x_{ij}}{\sum_{j=1}^{n} U_{ij}^{m}} \qquad (4)$$

Phase 3: This step computes the dissimilarities function, which determines the differences between the centroid and data points. Equation (6) is used to validate the threshold value, and if it is accurate, the step is finished; otherwise, it proceeds to step 3. It gauges how well the data clustering fits.

$$J(U, C_{1,}, C_2, \ldots\ldots\ldots\ldots C_c) = \sum_{i=1}^{c} j_i = \sum_{i=1}^{c}\sum_{j=1}^{n} U_{ij}^{m}\, d_{ij}^2 \qquad (5)$$

Where

$$d_{ij\|}^2 = \|CP_j - v_i\|^2 \qquad (6)$$

And $m \in [1, \infty]$ and m is a parameter that, when set to 2, controls how fuzzy the generated clusters will be. If

$$\| U^{(k+1)} - U^{(k)} \| < \in \qquad (7)$$

Equation (7) compares the threshold value to the difference between the value of the current classes and the upcoming classes of the membership function. If values are satisfied, we moved on to the subsequent phases.

Otherwise, repeat step 2 until the values are met and are set to 0.01.

Result and discussion

4.1 Result analysis from the SKM algorithm

Based on the number of times spent on a website and the types of website visited, the SKM algorithm clusters the data set to give the output seen below in Fig. 4

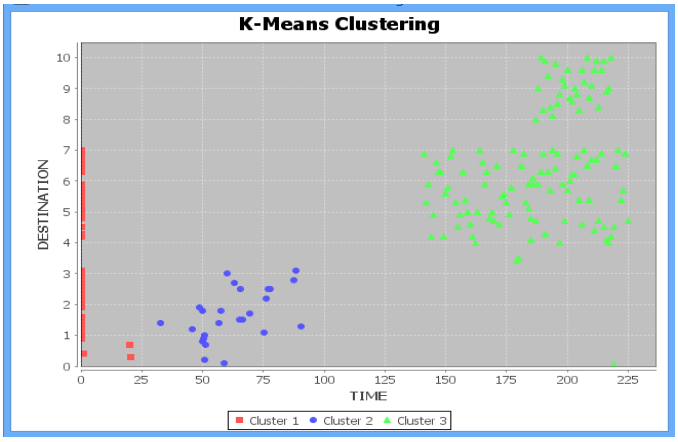

Fig. 4: Cluster graph

4.1.1 Discussion of result

From the diagram in fig. 4, we have three clusters generated. Cluster 1 showed that about 5% of the internet users visited websites considered good, cluster 2 showed 25% of the internet users visited websites were moderate and the last cluster showed a very high amount of

internet users visiting websites were considered bad, this means that they abused their access to the internet.

The result of this clustering showed that 38% of the data was grouped in class A; which means that 25% of the users on the network spend time on websites that were classified as moderate, that is website that cane considered not too bad, which means the users focus were on social media while 5% of the users were considered to make judicious used of the network, and 70% of the users were considered to have abused the used of the network. The bar graph showing the description of the percent usage is shown in Fig. 5.

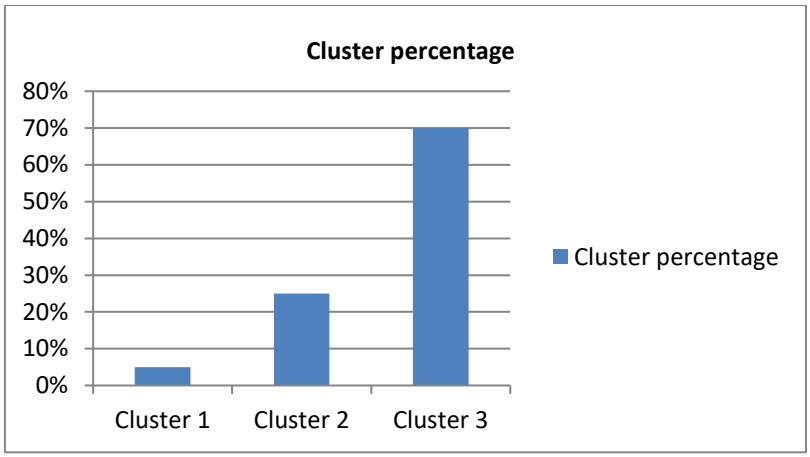

Fig. 5: Cluster percentage

4.2 Result analysis from the standard FCM algorithm

Fig. 6 displays the outcomes of the application of the conventional FCM clustering technique. The three clusters produced divided internet users into groups according to their level of behavior-based membership in the three clusters, such as legitimate, fair usage, and abuse. A particular data item of interest can be further researched by checking through its internet user profile history, including IP addresses, Apache log date, time, zone, section, Apache log file status code for the request, crawler, browser information, etc. to offer proof of abuse.

In Fig. 6, different clusters were created, and the degree to which each piece of data belonged to each cluster is shown by the membership value along the y axis. Internet users who share the same or comparable usage characteristics were grouped together based on the cluster that was created, and their level of membership, which showed how much a user belonged to the specific cluster, was also presented. This could be considered legal, ethical use, and abuse. The membership value only indicates how much each data point overlaps with the clusters.

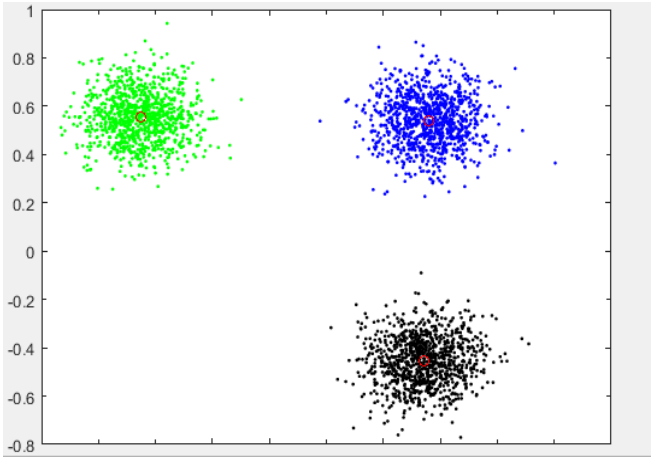

Fig. 6: Standard FCM

4.3 Evaluation of results

On the data set of internet users, the performance of the SKM and the conventional FCM were compared. Each of the two studies had 2150 instances and six attributes. Of the total data sets, 70% were utilized to train the system, while the remaining 30% were used to evaluate the system. The confusion matrices are relevant because the assessment metrics selected for this research are all based on the confusion matrices and the numbers inside of them. Experiment one was conducted using the conventional FCM and the confusion matrices given in Table 2. In Table 3, the performance result is shown. Sensitivity, accuracy, specificity, and execution time were taken into consideration when evaluating the performance. For further comparative study with other experiments conducted, the mean computation for the metrics for multiple studies was also computed. While Table 4 illustrates the second experiment's use of the SKM and its confusion matrix, Table 5 provides an explanation of the performance evaluation.

Table 2: Experiment one: confusion matrix using the standard FCM algorithm

Cluster No	True Positive (TP)	False Negative (FN)	False Positive (FP)	True Negative (TN)
1	213	35	26	69
2	169	33	29	153
3	148	28	9	100
Average	177	32	21	107

Table 3: Experiment one: performance analysis result using the standard FCM algorithm

Cluster No	Specificity(%)	Sensitivity(%)	Precision(%)	Accuracy(%)	Time(s)
1	85.87	86.18	94.22	86.09	0.32
2	84.06	83.33	90.31	83.59	0.32

3	91.74	84.09	94.27	87.018	0.32
Average	87.22	84.53	92.93	85.57	0.32

Table 4: Experiment two: confusion matrix using the SKM algorithm

Cluster No	True Positive (TP)	False Negative (FN)	False Positive (FP)	True Negative (TN)
1	208	30	27	81
2	164	37	29	150
3	158	229	10	91
Average	177	32	21	107

Table 5: Experiment two: performance analysis result using the SKM algorithm

Cluster No	Specificity(%)	Sensitivity(%)	Precision(%)	Accuracy(%)	Time(s)
1	85.87	86.18	94.22	86.09	0.13
2	84.06	83.33	90.31	83.59	0.12
3	91.74	84.09	94.27	87.01	0.16
Average	85.22	81.53	91.93	82.57	0.15

4.4 Analysis and comparison of the algorithm

Table 6 discussed the results gotten from the SKM and FCM algorithms used in the experiments. Five parameters were used to evaluate the performance of the algorithms. The standard FCM exhibited slight improvement over the SKM. The SKM had 85.2% specificity as against 87.2% for FCM; 81.5% sensitivity contrary to 84.5% for FCM; 91.9% precision compared to 92.9% for FCM; 82.5% accuracy against 85.6% FCM and a computational time of 0.15s and 0.32s for SKM and FCM respectively.

Table 6: Comparison analysis table of the algorithms

Algorithm	Specificity (%)	Sensitivity(%)	Precision (%)	Accuracy (%)	Time(s)
FCM	87.2	84.5	92.9	85.6	0.32
SKM	85.2	81.5	91.9	82.5	0.15

In relation to these results, regular FCM performed somewhat better than SKM with respect to all the metrics utilized in the evaluation apart from the execution time which can be trade off, given the system is supposed to be used for control purpose and at such, accuracy is

crucial. Hence, the high calculation time of the typical FCM could be exchanged for a better and accurate control system.

5. Conclusion

This research work studied and analyzed the use of clustering algorithm techniques in internet user profiling. The study further experiments the use of soft and hard clustering algorithm in internet profiling by performing a comparative analysis. Conclusively, from the result of the experiments carried out above, the standard FCM algorithm when compared with the SKM demonstrated that FCM performed better than SKM with respect to all the metrics used in the evaluation apart from the execution time, network administrators, using varying range of datasets and a more efficient clustering algorithm can now determine bandwidth usage for both productive ad unproductive business use and the same can be used for capacity planning.

Researchers in the areas of internet user profiling, bandwidth usage clustering and analysis, bandwidth optimization ect can further consider reducing the computational time of this algorithm. Lastly, future work in the area of improving on the standard algorithm in order to better optimized its performance is needed and results compared with the standard FCM.

References

1. S.S Banu, P. Uma, M.W. Ashfaque, Q.N. Naveed and S.S. Ali-Ahmed, “Data mining based soft computing skills towards preventio of cyber crimes on the web,” International Journal on Recent and Innovation Trends in Computing and Communication, vol. 3, issue 3, 2015, pp. 2321-8169. Retrivied from http://www.ijritcc.org/download/1427698079.pdf.
2. M. Ganesan and P. Mayilyahanan, “Cyber crime analysis in social media using data mining technique,” International Journal of Pure and Applied Mathematics, vol. 116, issue 22, 2017, pp. 1311-8080. Retrieved from https://acadpubl.eu/jsi/2017-116-13-22/articles/22/35.pdf.
3. V.S. Nath, “Crime patter detection using data mining,” Oracle Corporation, 2016, Retrieved from https://doi.org/10.1109/WI-IATW.2016.55.
4. A.M Neil, M. Elmogy and A.M. Riad, “Fuzzy crime investigation framework for tracking data theft basd on USB storaye,” International Journal of Computer Applications vol. 84, isseu 4, 2013, pp. 0975-8887. Retrieved from https://pdf.semanticscholar.org/6ed6/c1da68ded33e53787ab43eacfe582d8627f5.pdf.
5. I. Riadi, J.E. Isityano, A. Ashari and Subamar, “Log analysis techniques using clustering in network forensics,” International Journal of Computer Science and Information Security (IJCSIS), vol. 10, issue 7, 2010. Retrieved from http://eprints.uad.ac.id/6924/1/hasil-cek-similarity-0510088001-60020397-imamriadi-paper-c1.1.pdf.
6. R. Sahora, “Profiling a cyber crminal,” International Journal of Information and Communication Technology, vol. 4, isse 3, 2014, pp. 253-258. Retrieved from http://www.irphouse.com/ijict.htm. Retrieved from http://www.ijeit.com/vol%202/issue%203/IJEIT1412201209_37.pdf.

7. A. Sharma and S. Sharma, "An Intelligent analysis of web crime data minimg," International Journal of Engineering and Innovative Technology (IJEIT), vol. 2, issue 3, 2012, pp. 2277-3754.
8. S. Vashisht, M. Kaur and M. Singh, "Detecting cyber crime by analyzing user data," International Journal of Computer Technology and Application, vol. 3, issue 3, 2013, pp. 1029-1033. Retrieved from www.ijcta.com.
9. O. Maitanmi, S. Ogunlere, S. Ayinde and Y. Adekunle, Impact of cyber crimes on Nigeria economy," International Jurnal of Engineering and Science, vol.2, issue 4, 2013, pp. 2319-1813. Retrieved from http://www.theijes.com/papers/v2i4/part.%20(4)/H0244045051,pdf.
10. B. Metivier, "Cyber threat detection- 5 keys to log analysis success [infographic].
11. M. Zulfadhilah, Y. Prayudi, and I. Riadi, "Log classification using k-means clustering for identify internet user behaviors," International Journal of Advanced Computer Science and Application, vol. 7, issue 7, 2016,. Retrieved from http://www.theijes.com/papers/v2-i4/part.%20(4)/H024404551.pdf.
12. Z.S. Zubi, and A.A. Mahmmud, "Crime data analysis using data mining techniques to improve crimes prevention," International Journal of Computers, vol. 8, 2014, pp. 998-4308. Retrieved from www.naun.org/main/NAUN/computers/2014/a022007-096.pdf.
13. U. Mande, Y. Srinivas, and J.V.R. Murthy, "Witness based criminal identification using data mining techniques and new gaussian mixture model," International Journal of Modern Engineering Research (IJMER) www.ijmer.com, ISSN:2249-6645, vol. 2, issue 4, Jul-Aug, 2012, pp. 1507-1510.