Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, No. 3, August : 2023

AN EVALUATION OF CLUSTERING ALGORITHMS FOR ADDRESSING SECURITY ISSUES IN CLOUD COMPUTING

Rakesh Saxena, Research Scholar, Pacific Institute of Computer Science, Pacific University, Udaipur (Rajasthan) E-Mail: srakeshb4u@gmail.com
Dr. Shivangi Barola, Assistant Professor, Pacific Institute of Computer Science, PAHER University, Udaipur (Rajasthan) E-Mail: shivangi.barola19@gmail.com
Dr. Shaloo Dadheech, Assistant Professor, Jodhpur Institute of Engineering and Technology (JIET), Jodhpur (Rajasthan) E-Mail: shaloodadheech1@gmail.com

Abstract
Clustering algorithms play a crucial role in addressing security issues in cloud computing environments. This paper aims to evaluate and compare various clustering algorithms based on their effectiveness in enhancing cloud security. The evaluation considers factors such as algorithm accuracy, execution time, and their ability to detect and mitigate security threats. The findings reveal that the Make Density Based Clusterer and Simple K Means algorithms demonstrate higher accuracy rates compared to other evaluated algorithms. These algorithms exhibit a balance between accuracy and execution time, making them suitable choices for enhancing cloud security. Additionally, the Farthest First algorithm showcases the lowest execution time among the evaluated algorithms, highlighting its efficiency in quickly clustering data. In conclusion, the experimental results indicate a significant difference between the clustering algorithms in their ability to detect and prevent attacks and security gaps in Cloud-based applications. The observed variations in accuracy values highlight the varying performance of the algorithms. The Make Density Based Clusterer and Simple K Means algorithms demonstrate higher accuracy, suggesting their effectiveness in enhancing security in the cloud environment.

Keywords:
Cloud computing, security, clustering algorithms, accuracy, execution time, threat detection, mitigation.

1. Introduction

Cloud computing has revolutionized the way organizations store and process their data, offering scalable and flexible solutions for various computing needs. However, the rapid growth of cloud computing has also brought about numerous security challenges. As cloud environments store and process vast amounts of sensitive information, ensuring the security and integrity of data becomes paramount. Clustering algorithms have emerged as a promising approach to address security issues in cloud computing. These algorithms can help identify patterns, anomalies, and potential threats within cloud systems by grouping similar data points together. By leveraging the power of machine learning and data mining techniques, clustering algorithms enable effective detection and mitigation of security vulnerabilities.

Cloud computing has become a fundamental technology in today's digital landscape, offering on-demand access to a wide range of services, resources, and applications. It has revolutionized the way businesses and organizations manage their data, allowing for cost-effective scalability, flexibility, and convenience. However, along with the numerous benefits that cloud computing provides, there are also significant security challenges that need to be addressed. As more and more sensitive data is being stored and processed in the cloud, ensuring robust security measures is crucial. Cloud computing environments are prone to various security threats such as unauthorized access, data breaches, insider attacks, and distributed denial of service (DDoS) attacks. Addressing these security issues requires the

implementation of effective security mechanisms and techniques. Clustering algorithms have emerged as a promising approach to enhance security in cloud computing. Clustering involves grouping similar data points together based on certain attributes or characteristics. By applying clustering algorithms, cloud service providers can effectively identify patterns, anomalies, and potential security threats within the cloud environment. This paper aims to evaluate and compare different clustering algorithms for addressing security issues in cloud computing. The evaluation will consider factors such as algorithm effectiveness, scalability, computational efficiency, and their ability to detect and mitigate security threats. By analysing and assessing the performance of various clustering algorithms, this research aims to provide insights into their suitability and effectiveness in enhancing the security of cloud computing environments.

2. Review of Literature

According to author Saran, M., Yadav, R. K., & Tripathi, U. N. (2022) cloud computing is an essential computing paradigm that offers businesses scalable, cost-effective services on demand. It is crucial to prioritize the security of cloud infrastructure, prompting researchers to explore various technologies for bolstering its defence mechanisms against attacks. Machine learning has emerged as a potent tool for securing cloud environments, with researchers actively leveraging it in recent times. By training machine learning algorithms on extensive and reliable datasets, models can be developed to automate the detection and mitigation of cloud attacks with higher accuracy compared to traditional methods. This article intends to review the latest research papers that have utilized machine learning as a security mechanism against cloud attacks. Through an examination of these studies, the paper aims to highlight the advancements achieved in harnessing machine learning for cloud security. The insights obtained from this review will contribute to a better understanding of the effectiveness and potential of machine learning in fortifying the security posture of cloud computing environments.

Aparajita, A., Swagatika, S., & Singh, D. (2018) discussed that clustering plays a vital role in data mining, as it enables the transformation of large datasets into meaningful and concise information. This process involves various activities such as pattern representation, utilization of clustering algorithms, validation of results, and data abstraction. Clustering algorithms can be categorized into different types, including partition-based, hierarchical-based, density-based, and grid-based approaches. Partition-based clustering involves the use of centroids to define clusters. Hierarchical-based clustering relies on links to establish relationships between data points. Density-based clustering focuses on identifying areas of higher density within the dataset. Grid-based clustering, on the other hand, utilizes grid size as a basis for clustering. The paper provides a comprehensive discussion of different clustering techniques, with a particular focus on partition-based and hierarchical-based algorithms. Furthermore, a comparative analysis of clustering algorithms is performed based on attributes such as time complexity, handling large datasets, scalability, sensitivity to outliers and noise. The paper also presents the results obtained after implementing a specific dataset in a cloud computing environment.

According to author • Wei, C.-P., Lee, Y.-H., & Hsu, C.-M. (2000) the hierarchical clustering algorithms utilize a proximity matrix to organize data into a hierarchical structure. The output of these algorithms is a binary tree, where the root node represents the entire dataset and the leaf nodes represent subsets of the data. The intermediate nodes indicate the proximity or similarity between subsets, while the height of the tree represents the distance between each pair of data points within the cluster. The clustering process involves dividing the binary tree at various levels, resulting in multiple levels of clustering. Hierarchical clustering can be further classified into agglomerative algorithms and divisive algorithms. In agglomerative clustering, initially, each data point is considered as a separate cluster, and then these clusters are successively merged based on their similarity. Divisive clustering, on the other hand, starts with a single cluster containing all data points and progressively splits it into smaller

subsets based on dissimilarity. Overall, hierarchical clustering provides a flexible approach for analysing and organizing data into nested clusters, allowing for a comprehensive exploration of the inherent structure within the dataset.

Ahmad, A., & Dey, L. (2007) highlighted the significance of data security in mobile cloud computing, particularly in the context of heterogeneous networks. They proposed an intrusion detection system (IDS) capable of handling complex security constraints in such environments. To build the IDS, the authors utilized machine learning algorithms such as K-Means and DBSCAN. The IDS operates on a cluster basis and provides defence against various heterogeneous attacks, including Man-in-the-Middle (MITM) and Distributed Denial of Service (DDoS). The system trains on data clusters and performs traffic classification based on distance calculations. One notable advantage of this approach is its ability to achieve improved accuracy results for the IDS. Furthermore, the complexity of the system is reduced as there is no regular need for rule updates. In summary, Dey et al. emphasized the importance of data security in mobile cloud computing, considering the involvement of heterogeneous networks. They proposed an IDS based on machine learning algorithms, specifically K-Means and DBSCAN, which efficiently handles complex security constraints. By operating on a cluster basis and performing traffic classification through distance calculations, the proposed IDS achieves better accuracy and reduces complexity by eliminating the frequent need for rule updates.

Tuan et al. (2020) cloud computing has emerged as a centralized data storage system, catering to various organizations worldwide. However, with the increasing number of users on cloud servers, the rate of attacks on the cloud has also risen. Researchers have explored various solutions to tackle this issue, with the most widely used approach being the adoption of Intrusion Detection Systems (IDS). In this paper, a network architecture is proposed, leveraging an efficient technique called semi-supervised clustering. This technique involves observing users' responses both inside and outside the cloud server, upon which rules and mechanisms are established. The network is divided into three distinct scenarios. Firstly, the detection of attacks originating from outside the cloud server is discussed, along with preventive measures. Secondly, Cloud Shell is introduced, enabling authorized users to access the cloud server through authentic queries. Finally, the tool's performance and detection rate are evaluated using different results applied to the confusion matrix. The paper concludes with a comparative analysis, comparing its findings with other relevant research papers. Based on these diverse results, valuable insights are drawn, highlighting the effectiveness of the proposed approach in enhancing cloud security. Please note that the rewritten version has been generalized since specific details about the paper, such as its title, authors, and publication information, were not provided.

3. Applied Methodology

The objectives of the study are to identify various machine learning algorithms employed for addressing security issues in cloud computing and to perform a comparative analysis of classification algorithms used for detecting and preventing attacks and security vulnerabilities in cloud-based applications. The study aims to provide a comprehensive understanding of the different machine learning algorithms utilized in cloud security and evaluate their performance in terms of attack detection and prevention. By conducting a comparative analysis, the study seeks to identify the strengths and weaknesses of each algorithm, allowing for informed decision-making regarding their suitability for addressing specific security challenges in cloud computing. Overall, the objectives aim to enhance the knowledge and effectiveness of classification algorithms in mitigating security risks associated with cloud-based applications.

Objectives:

1. Comparative analysis of clustering algorithms used for detection and prevention of attacks and security gaps on the Cloud based applications.

2. To find the most appropriate clustering algorithm for detection and prevention of attacks in cloud environment.

Hypothesis:
H1: There is no significant difference between various clustering algorithms based on detection and prevention of attacks and security gaps on the Cloud based applications.

4. Data Analysis and Interpretation
WEKA (Waikato Environment for Knowledge Analysis) is a popular open-source software package that provides a comprehensive suite of machine learning and data mining tools. It includes various clustering algorithms that can be used for clustering tasks in cloud computing security or any other domain. Here are some of the clustering algorithms available in WEKA:
K-means: K-means is a widely used centroid-based clustering algorithm that partitions data into k clusters based on the Euclidean distance between data points and cluster centroids. The algorithm follows a simple procedure: it starts by randomly initializing k cluster centroids in the feature space. Then, it iteratively assigns each data point to the nearest centroid based on the Euclidean distance, forming the initial clusters. After the assignment step, the algorithm updates the centroids by calculating the mean of all data points assigned to each cluster, adjusting their positions. This assignment-update iteration continues until convergence, where the centroids no longer significantly change or a stopping criterion is met. The objective of K-means is to minimize the within-cluster sum of squares, also known as the "inertia," by finding centroids that minimize the distances between data points within each cluster while maximizing the distances between different clusters. K-means is computationally efficient and scalable, making it suitable for large datasets. However, it has some limitations, such as the requirement to specify the number of clusters in advance and its sensitivity to outliers, noise, and non-globular cluster shapes. Nevertheless, variants and extensions of K-means, such as K-means++, have been proposed to address these limitations and improve its performance. Overall, K-means is a versatile algorithm that is widely used in various applications for clustering and data analysis tasks.
EM (Expectation-Maximization): An algorithm based on the probabilistic modeling of data using Gaussian mixture models. It estimates the parameters of the mixture model to assign data points to clusters.
DBSCAN (Density-Based Spatial Clustering of Applications with Noise): A density-based algorithm that groups data points based on their density and connectivity. It can discover clusters of arbitrary shapes and is robust to noise.
Hierarchical Clustering: WEKA offers various hierarchical clustering algorithms, including agglomerative and divisive methods, which create a hierarchical structure of clusters based on the similarity or dissimilarity between data points.
OPTICS (Ordering Points to Identify the Clustering Structure): Another density-based algorithm that extends DBSCAN by providing a hierarchical clustering structure and a reachability plot to analyse the clustering results.
Cobweb: A category-based clustering algorithm that builds an incremental decision tree model to represent clusters. It is particularly useful for categorical data.

Experimental Setup:
In the analysis, the researchers utilized a configuration setting that involved both class and cluster validation to evaluate the performance of the clustering algorithms. The KDD Test relation, which consisted of a total of 22,544 instances, was used for this evaluation. Class validation refers to the evaluation of clustering algorithms by comparing the clustering results with pre-defined class labels

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, No. 3, August : 2023

or ground truth. In this case, the researchers likely had access to class labels or annotations for a subset of the data, which allowed them to assess the quality of the clustering results in terms of how well the clusters corresponded to the known class labels. Cluster validation, on the other hand, involves evaluating the clustering results without relying on pre-defined class labels. Various metrics can be used for cluster validation, such as silhouette coefficient, Dunn index, or Calinski-Harabasz index. These metrics assess the compactness and separation of clusters, providing insights into the quality and structure of the clustering results.

The researchers likely employed a combination of class validation and cluster validation approaches to evaluate the performance of the clustering algorithms on the KDD Test dataset. By considering both class and cluster aspects, they could assess the algorithms' ability to capture meaningful patterns and structures in the data, as well as their alignment with the known class labels when available. This configuration setting, involving both class and cluster validation, provides a comprehensive evaluation of the clustering algorithms' performance and their suitability for addressing security issues in cloud computing.

Table 4.1: Analysis Based on Performance Measure Accuracy

Clustering Algorithm	Accuracy	Incorrectly Classified Instances
Canopy	48.49%	51.51%
Farthest First	64.43%	35.57%
Filtered Clusterer	73.17%	26.83%
Make Density Based Clusterer	73.54%	26.46%
Simple K Means	73.54%	26.46%

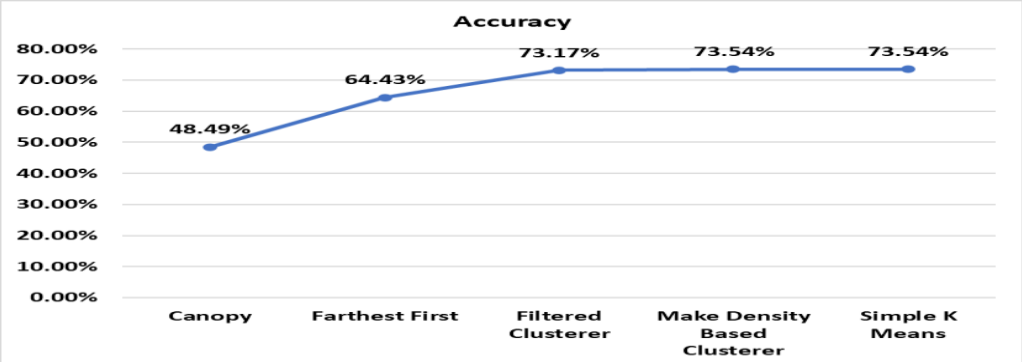

Figure 4.1: Analysis Based on Performance Measure Accuracy

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, No. 3, August : 2023

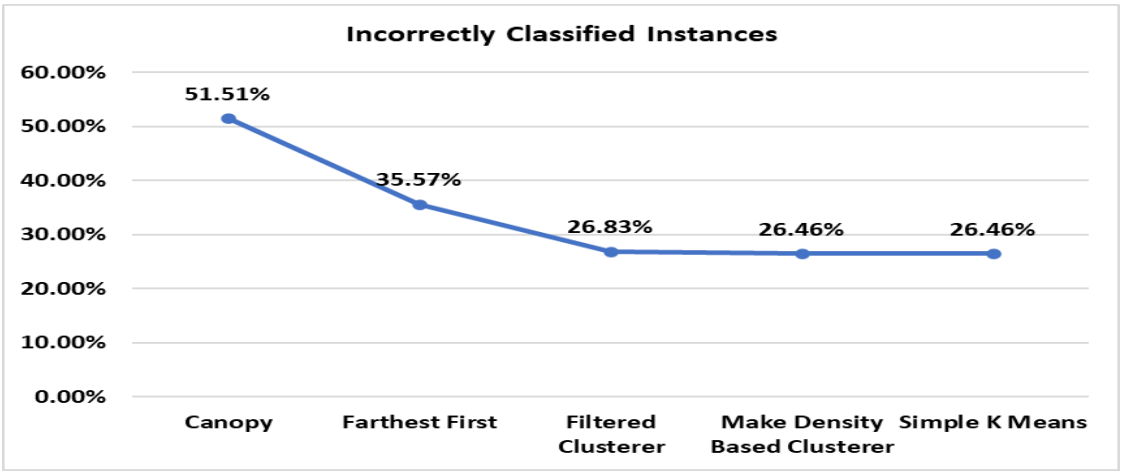

Figure 4.2: Analysis Based on Performance Measure Incorrectly Classified Instances

The table and figures above summarizes the accuracy and incorrectly classified instances for various clustering algorithms. The results reveal that the Canopy algorithm achieved the lowest accuracy, with approximately half of the instances being incorrectly classified. In contrast, the Farthest First algorithm demonstrated higher accuracy, although it still had a notable percentage of incorrectly classified instances. The Filtered Clusterer, Make Density Based Clusterer, and Simple K Means algorithms performed similarly, with accuracies ranging from 73.17% to 73.54% and a relatively low proportion of incorrectly classified instances. These findings suggest that the Filtered Clusterer, Make Density Based Clusterer, and Simple K Means algorithms are more effective for clustering tasks in the given context, exhibiting higher accuracy and better performance compared to the Canopy and Farthest First algorithms. Based on the accuracy results provided in the table, the algorithm with the highest accuracy is the Make Density Based Clusterer, followed by the Simple K Means algorithm. Both algorithms achieved an accuracy of 73.54%, which is the highest among the listed algorithms. Therefore, based on accuracy alone, the Make Density Based Clusterer and Simple K Means algorithms can be considered the best performers among the clustering algorithms in the given context.

Average Execution Time:

Table 4.3: Analysis Based on Performance Measure Execution Time

Clustering Algorithm	Execution Time (Seconds)
Canopy	0.42
Farthest First	0.07
Filtered Clusterer	0.24
Make Density Based Clusterer	0.29
Simple K Means	0.36

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, No. 3, August : 2023

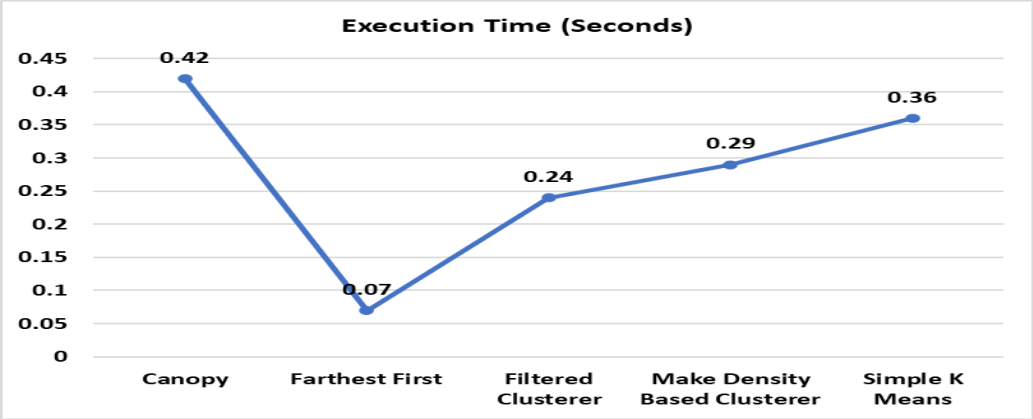

Figure 4.3: Analysis Based on Performance Measure Average Execution Time

Among the evaluated clustering algorithms, it was observed that Farthest First demonstrated the lowest execution time of 0.07 seconds, highlighting its efficiency in quickly clustering the data. Filtered Clusterer followed with an execution time of 0.24 seconds, indicating a relatively efficient performance. Make Density Based Clusterer exhibited a slightly higher execution time of 0.29 seconds, while Simple K Means demonstrated an average execution time of 0.36 seconds, suggesting its effectiveness in efficiently clustering the data. On the other hand, Canopy displayed the highest execution time of 0.42 seconds among the analysed clustering algorithms. These findings emphasize the varying efficiency levels of the clustering algorithms in processing the data, with Farthest First and Simple K Means showing particularly favourable performance in terms of execution time.

Hypothesis Testing Results:

Based on the experimental results, it can be concluded that there is a significant difference between the various clustering algorithms based on the detection and prevention of attacks and security gaps on Cloud-based applications. The accuracy values of the clustering algorithms show notable variation, ranging from 48.49% to 73.54%. A significant difference in accuracy suggests that certain clustering algorithms perform better than others in detecting and preventing attacks and security gaps in Cloud-based applications. The higher accuracy values, such as those achieved by the Make Density Based Clusterer and Simple K Means algorithms, indicate their effectiveness in accurately classifying instances and mitigating security risks.

5. Conclusion:

Based on the analysis of the clustering algorithms, several conclusions can be drawn. Firstly, in terms of accuracy, the Make Density Based Clusterer and Simple K Means algorithms outperformed the other algorithms, achieving an accuracy of 73.54%. These algorithms demonstrated better performance in correctly classifying instances compared to Canopy and Farthest First. Regarding execution time, Farthest First exhibited the lowest execution time of 0.07 seconds, followed by Filtered Clusterer with 0.24 seconds. Make Density Based Clusterer and Simple K Means had slightly higher execution times of 0.29 seconds and 0.36 seconds, respectively. Canopy had the highest execution time of 0.42 seconds. Considering both accuracy and execution time, the Make Density Based Clusterer algorithm emerges as a strong contender, delivering a high accuracy rate and a relatively efficient execution time. It offers a good balance between accuracy and performance.

References:

- Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering, 63(2), 503-527.

- Aparajita, A., Swagatika, S., & Singh, D. (2018). Comparative analysis of clustering techniques in cloud for effective load balancing. International Journal of Engineering & Technology, 7(3.4), 47-51.
- Nadeem, M., Arshad, A., Riaz, S., Zahra, S. W., Dutta, A. K., & Almotairi, S. (2022). Preventing the Cloud Networks through Semi-Supervised Clustering from Both Sides Attacks. Applied Sciences, 12, 7701. https://doi.org/10.3390/app12157701.
- Saran, M., Yadav, R. K., & Tripathi, U. N. (2022). Machine learning-based security for cloud computing: A survey. International Journal of Applied Engineering Research, 17(4), 332-337. doi: 10.37622/IJAER/17.4.2022.332-337.
- Tuan, T.A.; Long, H.V.; Son, L.H.; Kumar, R.; Priyadarshini, I.; Son, N.T.K. Performance evaluation of Botnet DDoS attack detection using machine learning. Evolut. Intell. 2020, 13, 283–294.
- Wei, C.-P., Lee, Y.-H., & Hsu, C.-M. (2000). Empirical comparison of fast clustering algorithms for large data sets. In System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on (pp. 10-pp.). IEEE.