International Information and Engineering Technology Association
Advancing the World of Information and Engineering

International Journal of Safety and Security Engineering
Vol. 15, No. 4, April, 2025, pp. 777-785
Journal homepage: http://iieta.org/journals/ijsse

Comparative Analysis of Clustering Algorithms for Financial Fraud Detection

Rajashekhar Rao K.*, Venkata Naresh Mandhala

Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram 522302, India

Corresponding Author Email: krsraohyd@gmail.com

https://doi.org/10.18280/ijsse.150414

Received: 4 January 2025
Revised: 16 April 2025
Accepted: 23 April 2025
Available online: 30 April 2025

Keywords:
financial fraud detection, K-means, DBSCAN, OPTICS, clustering evaluation

ABSTRACT

The detection of financial fraud has become an increasingly critical concern in today's data-driven economy, necessitating the development and application of robust analytical methods. This research undertakes a comparative analysis of various clustering techniques, specifically partitioning methods such as k-means, k-medoids, CLARA, CLARANS, BIRCH, and density-based algorithms like DBSCAN and OPTICS, alongside hierarchical clustering methods. By evaluating these algorithms, the study aims to identify the most effective method for un- covering patterns indicative of fraudulent activities within financial datasets. To gauge the performance of these clustering techniques, several evaluation metrics will be employed, including the Rand Index, Adjusted Rand Index (ARI), silhouette coefficient, and Davies-Bouldin Index (DBI). The Rand Index serves as a foundational measure for assessing clustering efficacy by quantifying agreement between predicted and true clusters. The ARI enhances this evaluation by accounting for chance agreements, thereby providing a more nuanced understanding of clustering performance. The silhouette coefficient offers insights into the cohesion and separation of clusters, while the DBI assesses cluster quality by evaluating intra-cluster and inter- cluster distances. This comprehensive analysis not only aims to determine the optimal clustering method for financial fraud detection but also seeks to contribute to the broader field of unsupervised machine learning. By systematically comparing the strengths and weaknesses of various clustering approaches, this research endeavours to provide valuable insights and guidelines for practitioners in the finance sector, enhancing their ability to detect and mitigate fraudulent activities effectively.

1. INTRODUCTION

Nowadays, many financial companies are manipulating their financial reports and committing fraud to gain their financial benefits. So, there is a severe need to eradicate such crimes, and so fraud detection has become much more important. Many methods have been identified to detect fraud. One of the methods is to use machine learning and develop tools for identification and fraud detection. We have a lot of algorithms using which tools have been developed. These algorithms are used to learn or identify whether the financial statements are fraudulent or non-fraudulent. These algorithms can be divided into 2 categories according to their learning process. The first is supervised learning, and the second is unsupervised learning.

Supervised learning is the process of learning on a set of labeled data, whereas unsupervised learning is the process of learning on a set of unlabeled data. Basically, this learning method is used for finding patterns and relationships in the data. In this paper, we are going to use some of the unsupervised learning algorithms that are used to check their capability of identifying the patterns that lead to fraud detection.

Clustering comprises a wide range of methodologies aimed at identifying subgroups or clusters by characterizing objects within a dataset, ensuring that objects within the same group exhibit similarity while differing from those in other groups. The fundamental principle of clustering is that data within a cluster should exhibit high similarity to one another while demonstrating significant dissimilarity from data in other clusters. Various clustering approaches exist, including partitioning methods, hierarchical methods, and density-based methods.

The K-Means algorithm [1] is a very effective unsupervised learning method that adeptly partitions datasets into discrete clusters based on the inherent similarities among data points. This approach functions on a centroid-based paradigm, wherein each cluster is linked to a central point referred to as the centroid. The primary aim of K-Means is to enhance the clustering of data points so that similar or proximate points are aggregated, resulting in cohesive clusters. The procedure entails the iterative allocation of data points to the closest centroid and the subsequent recalibration of the centroid's location based on the average of the allotted points, finally enhancing the clusters until convergence is reached.

The K-Medoids algorithm [2] is a strong clustering technique that resembles K-Means but differs fundamentally by employing medoids rather than centroids. A medoid is characterized as the most centrally situated point inside a cluster, rendering K-Medoids very proficient in managing datasets containing outliers. In contrast to K-Means, which is significantly affected by outliers, K-Medoids prioritizes the minimization of the aggregate distance between the medoid and all other points in the cluster, thus enhancing the clustering's robustness against data anomalies.

K-Means and K-Medoids are classified as partitioning methods, a vital group of clustering approaches in data analysis. These methods are essential for analyzing and comprehending intricate datasets, allowing researchers and practitioners to reveal concealed patterns and correlations within the data. Utilizing the advantages of these algorithms enables the attainment of significant insights and facilitates informed decision-making across many applications, including market segmentation and image processing.

The agglomerative clustering technique [3] is distinguished for its effective bottom-up methodology in clustering. It commences with each individual data point regarded as a separate cluster. As the algorithm advances, it methodically consolidates these clusters according to their intrinsic commonalities, thereby creating larger and more significant groupings. The iterative merging procedure persists until it fulfills a user- defined condition or until all clusters amalgamate into a singular cohesive cluster. A dendrogram, resembling a tree structure, functions as a visual representation of the hierarchical clustering process. This graphic represents individual data points as leaf nodes and clusters as root nodes, offering a clear comprehension of the relationships and hierarchies within the data.

CLARA [4], an acronym for "Clustering Large Applications," represents a significant advancement in the k-medoids (PAM) methodology. It is intended to facilitate the management of datasets containing numerous objects, typically exceeding several thousand observations. This novel method proficiently tackles the issues related to computational duration and memory capacity through the utilization of a strategic sampling methodology. By choosing representative samples from the dataset, CLARA can execute clustering without processing the complete dataset simultaneously, thus optimizing the clustering procedure and improving efficiency.

CLARANS [5], an acronym for Clustering Large Applications based on Randomized Search, is a unique partitioning technique for clustering, especially beneficial in geographical data mining. This method is proficient in revealing patterns and relationships inherent in spatial datasets, which may encompass distance-related, directionally related, or topological information, such as data depicted on a road map. CLARANS employs advanced spatial data mining algorithms to identify and categorize patterns, hence enhancing understanding of the geographical relationships among data points. BIRCH [6], or Balanced Iterative Reducing and Clustering utilizing Hierarchies, is an innovative multiphase clustering technique that functions through two essential phases: the formation of the Clustering Feature (CF) Tree and the implementation of global clustering. The initial phase focuses on compressing extensive datasets into more compact, denser areas known as CF vectors. This approach accomplishes substantial data reduction while encapsulating three critical summary statistics: count (N), linear sum (LS), and squared sum (SS), to accurately depict densely packed sub-clusters. The resultant CF Tree establishes a multilevel hierarchy that effectively consolidates smaller clusters into bigger, more complete entities, employing the notion of vector addition to optimize clustering efficiency.

DBSCAN [7], which stands for Density-Based Spatial Clustering of Applications with Noise, is an innovative technique that employs the distances among data points to create clusters. It functions according to two essential parameters: minPts and ε. The minPts option specifies the lowest number of points necessary to form a valid cluster, whilst the ε parameter establishes the maximum distance within which points are regarded as belonging to the same cluster. The procedure commences by randomly selecting a point from the dataset and finding all surrounding points within the ε distance. If the quantity of nearby points meets or surpasses the minPts level, they are collectively considered part of the same cluster. This procedure is recursively implemented for all sites identified within the cluster. If the quantity of nearby points is insufficient to meet the minPts criterion, the algorithm designates that point as an outlier. The classification procedure persists until each point in the dataset has been assessed and categorized, culminating the algorithm's function after all points have been designated as either belonging to a cluster or classified as outliers. OPTICS, an acronym for Ordering Points to Identify the Clustering Structure, is a robust density-based clustering method that resembles DBSCAN yet offers the exceptional capability to identify clusters of diverse densities and geometries. This capacity is especially beneficial for identifying clusters with varying densities in large, high-dimensional datasets. The core premise of OPTICS [8] is to clarify the clustering structure of a dataset by recognizing density-connected points. The algorithm carefully creates a density- based representation of the data using an ordered list of points referred to as the reachability plot. Every item in this list is linked to a reachability distance, measuring the accessibility of that item from other places in the collection. Points with comparable reachability distances are likely to be part of the same cluster, hence improving the overall efficacy and precision of the clustering process.

The objective of this work is to determine the most accurate clustering method, utilizing a dataset that includes class values solely for the purpose of assessing algorithmic accuracy. Evaluation measures are employed to examine the quality of clustering findings. These measures assess the internal coherence of clusters and the inter-cluster separation. Prevalent evaluation criteria comprise the Rand Index, adjusted Rand Index, silhouette score, Davies-Bouldin index, among others. We employed these methods to ascertain the efficiency of the algorithms presented in this research.

A confusion matrix [9] is an effective instrument that offers a detailed overview of a machine learning model's efficacy, especially regarding classification tasks. This matrix visually represents the counts of true positives, true negatives, false positives, and false negatives generated by the model on a specified test dataset. Through the analysis of these values, practitioners can obtain insights into the model's performance, pinpointing specific areas of proficiency or deficiency. The confusion matrix is particularly advantageous for assessing classification models that seek to allocate categorical labels to input examples, facilitating a detailed comprehension of the model's predicted performance. The Rand Index is a vital indicator for evaluating the effectiveness of clustering methods. Clustering, an unsupervised machine learning technique, aims to aggregate analogous data points into unified clusters. The Rand Index assesses the quality of clusters by evaluating the concordances and discordances between pairs of data points in the predicted clusters relative to those in the actual clusters. The Rand Index [10] produces a singular numerical score that represents the ratio of agreements, offering a definitive measure of the clustering algorithm's efficacy. This metric is essential for assessing the similarity across diverse clustering results, enabling researchers to evaluate the efficacy of different approaches or algorithms utilized in the analysis. The Adjusted Rand Index (ARI) [10] refines the Rand Index by incorporating a correction for chance agreements in the comparison of clusterings. This modification is especially crucial in situations when the quantity of clusters or their dimensions may occur solely by random chance, as the conventional Rand Index may yield deceptive outcomes in these circumstances. The Adjusted Rand Index computes the Rand Index while considering the anticipated degree of similarity between two random clusterings of an identical dataset. This enhances the ARI as a more robust instrument for clustering analysis, enabling a more precise evaluation of the alignment of clusters generated by various methods with one other or with reference clustering, also known as the ground truth. Another essential statistic in the evaluation of clustering quality is the silhouette coefficient. This metric analyzes the suitability of each data point's assignment to its respective cluster by measuring two essential aspects: cohesion and separation. Cohesion relates to how closely associated a data point is to other points within the same cluster, while separation denotes how distinctly a point is positioned compared to points in other clusters. The silhouette coefficient varies from -1 to 1, with a value approaching 1 indicating effective clustering, a value around 0 implying possible cluster overlap, and a value close to -1 suggesting potential misclassification of the point. A superior silhouette score indicates effective clustering, characterized by distinct boundaries between clusters and cohesive groups within them. A diminished score may indicate errors, such as incorrect point assignments to clusters or overlapping clusters. The Davies-Bouldin Index (DBI) [11] is a crucial indicator for assessing the validity of clustering solutions. This index computes the average similarity between each cluster and its most analogous cluster, based on two essential elements: within-cluster distance and between-cluster distance. The within-cluster distance represents the mean distance from data points to their corresponding cluster centroid, whereas the inter-cluster distance quantifies the distance between the centroids of distinct clusters. The DBI is calculated as the mean of the greatest similarity ratios for each cluster, with a larger similarity ratio signifying inadequate separation and delineation of clusters. A reduced DBI value signifies a more effective clustering solution. To utilize the DBI for assessing clustering results, one may calculate the index across different clustering techniques or parameter configurations and juxtapose the findings. The clustering approach with the lowest Davies-Bouldin Index (DBI) is deemed the most successful. The DBI can aid in identifying the ideal number of clusters for a dataset. By graphing DBI values against varying cluster counts, analysts can discern the "elbow point," characterized by a substantial decline in DBI, succeeded by a plateau. This point represents the ideal equilibrium between intra-cluster similarity and inter-cluster dissimilarity, assisting practitioners in determining the suitable number of clusters for their data analysis.

A confusion matrix is a matrix that encapsulates the performance of a machine learning model on a certain set of test data. It serves to illustrate the quantity of correct and incorrect examples depending on the model's predictions. It is frequently utilized to assess the efficacy of classification algorithms designed to predict a categorical label for each input occurrence.

Accuracy serves as a metric to evaluate the model's performance. It is the proportion of accurate occurrences to the total instances.

The Rand Index is a statistic employed to assess the efficacy of a clustering method. Clustering is an unsupervised machine learning technique employed to aggregate analogous data into a singular cluster, with the Rand Index indicating the efficacy of the clustering process. It evaluates the accumulation of data point pairs within the expected cluster compared to the actual cluster. The Rand Index yields a singular score reflecting the degree of concordance between the two groups.

The Rand Index is a metric used to assess the similarity between two distinct data clusterings. It evaluates the degree of concordance between the clusters generated by two distinct methods or algorithms.

However, the Rand Index does not account for the potential of accidental agreements between the two parties. The Adjusted Rand Index (ARI) is frequently employed to address unpredictability. The Adjusted Rand Index (ARI) enhances the Rand index to produce a statistic that can take on negative values when the agreement is below what random chance would predict, but a value of 1 indicates total agreement.

The Adjusted Rand Index (ARI) is an enhancement of the Rand index (RI) that accounts for randomness in assessing the similarity of two data clusterings. This measure is utilized in clustering analysis to evaluate the concordance of clusters generated by various approaches or algorithms with one another or with a reference clustering (ground truth).

In scenarios where the quantity or dimensions of clusters in the dataset may arise by random chance, the Rand Index may produce deceptive outcomes. The Adjusted Rand Index mitigates this drawback by accounting for random agreements. It calculates the Rand Index, considering the anticipated resemblance between two random clusterings of identical data.

The silhouette coefficient measures the extent to which each data point corresponds with its assigned cluster. It incorporates data concerning both cohesion (the closeness of a data point to other points within its cluster) and separation (the distance of a data point from points in different clusters).

The coefficient ranges from -1 to 1, where a value close to 1 indicates well-clustered data points, a value near 0 suggests overlapping clusters, and a value near -1 signifies misclassified data points.

An elevated silhouette score [12] indicates that the data points are well-clustered, demonstrating clear separation across clusters and robust cohesiveness within each cluster. A reduced silhouette score suggests that the grouping may lack precision, marked by overlapping groups or points improperly allocated to their respective clusters.

The Davies-Bouldin index (DBI) is a statistic for cluster validity that measures the average similarity between each cluster and its nearest counterpart. Similarity is assessed based on two criteria: intra-cluster distance and inter-cluster distance. The within-cluster distance is the average distance of data points in a cluster to its centroid. The between-cluster distance denotes the separation between the centroids of two separate clusters.

The DBI denotes the average of the maximum similarity ratios for each cluster. The similarity ratio is determined by dividing the sum of within-cluster distances by the distance between clusters. A high similarity ratio signifies that the clusters exhibit insufficient separation or definition. A diminished DBI indicates a more optimal clustering solution.

What is the procedure for employing the DBI to evaluate your clustering solution? One can calculate the DBI for different clustering methodologies or parameters and conduct a comparison analysis. The alternative with the minimal DBI is the most advantageous. The DBI can be employed to determine the optimal number of clusters for your dataset. One can plot the DBI values for different cluster metrics and discern the elbow point, where the DBI exhibits a significant decrease followed by a plateau. This is the optimal number of clusters that balance intra-cluster similarity and inter-cluster dissimilarity.

The above algorithms are being used in this paper as they are rooted in their ability to perform unsupervised learning effectively, allowing them to discover hidden patterns and natural groupings in data without the need for labeled examples. These algorithms are highly versatile, working across different domains and data types, and can handle multidimensional datasets, making them useful for exploratory data analysis, segmentation, and pattern recognition. They help in simplifying complex data by organizing it into clusters, which can improve interpretability and support downstream tasks like classification or anomaly detection. Additionally, many of these methods are available in popular machine learning libraries, making them accessible and practical for real-world applications. Together, they offer a balance of efficiency, scalability, robustness to noise, and adaptability to various data structures.

2. LITERATURE REVIEW

Huang et al. [13] presented a novel dual GHSOM (Growing Hierarchical Self-Organizing Map) approach for detecting fraudulent financial reporting (FFR) and extracting relevant features from financial data. They utilize a dataset of 762 financial statements from 144 publicly traded companies in Taiwan, identifying 72 fraudulent and 72 non-fraudulent samples. The study demonstrates that the topological patterns of FFR can effectively distinguish between fraudulent and non-fraudulent samples, achieving promising results in classification and feature extraction. The proposed method shows effectiveness with Type I and Type II errors below 20%, indicating reliable decision support for identifying potential FFR categories.

The study conducted experiments using real fraudulent financial reporting (FFR) statements to validate the proposed dual GHSOM approach, demonstrating its effectiveness in detecting FFR and extracting relevant features. The experimental results indicated that the topological patterns of FFR follow a non-fraud-central spatial relationship, suggesting the potential of these patterns for effective FFR detection. The classification results of the dual GHSOM approach were compared with other methods, showing its superiority in identifying fraudulent samples and extracting salient characteristics of fraud behaviors.

Li et al. [14] presented an integrated approach for automatic cluster detection, optimization, and interpretation in financial data, addressing the complexity of human behaviors and data distributions. They introduce a new cluster quality evaluation criterion that guides base clustering algorithms to detect hyper-ellipsoidal clusters adaptively. The proposed method includes a revised support vector data description model to refine cluster centroids and scopes, enhancing interpretability. Experiments on ten financial datasets demonstrate the algorithm's efficiency in identifying a reasonable number of clusters suitable for financial mining tasks.

The study proposed a new clustering approach called Ad-Ellip, which effectively detects hyperellipsoidal clusters in financial data. Ada-Ellip demonstrated superior performance in terms of cluster quality evaluation, achieving reasonable cluster numbers and tight clusters with high data point similarities. The algorithm was notably faster than traditional methods, making it suitable for large-scale financial datasets. Ten financial benchmark datasets were used to show that Ada-Ellip is good at automatically finding and interpreting clusters, which makes it useful for tasks like fraud detection and credit evaluation.

Thiprungsri and Vasarhelyi [15] explored the use of cluster analysis for anomaly detection in accounting, specifically in auditing group life insurance claims. They identify discrepancies by grouping claims with similar characteristics and flagging small- population clusters for further investigation. The dataset consists of 208 attributes related to claims, with 169 identified as possible anomalies based on cluster membership probabilities. The study emphasizes the importance of domain knowledge in evaluating clustering results and suggests that cluster analysis can enhance fraud detection techniques in auditing.

The study identified a total of 169 claims as possible anomalies based on cluster-based outliers. It was determined that 568 claims had a probability of less than 0.6 of belonging to their assigned clusters, marking them as potential anomalies. The clustering procedure utilized simple K-means, resulting in eight clusters from a dataset of 40,080 claims paid in the first quarter of 2009. The analysis revealed that clusters with small populations exhibited unusual characteristics, such as high interest-to-beneficiary payment percentages and extended periods between death dates and payment dates.

Tatusch et al. [16] presented a novel approach to detecting financial restatements using a modified, dynamic version of the DBSCAN clustering algorithm. They analyze data from 9300 companies over a 20-year period (1998–2017) to identify restatements based on four definitions. The modified DBSCAN algorithm excels in precision, achieving over 50% accuracy in classifying firm-years as restatement or non-restatement years. The study highlights the importance of data processing methods over sheer data volume in improving detection efficiency. The paper presents a modified version of the DBSCAN clustering algorithm, achieving over 50% accuracy with just two or three features. The model demonstrates high efficiency in detecting restatement years compared to non-restatement years. Precision values vary, with the best performance noted at 65.6% for original data and 56.7% for processed features. The results indicate that non-restatements are generally better identified than restatements, highlighting the model's strengths in precision over accuracy. Overall, the approach outperforms prior methods in identifying financial restatements.

Herman et al. [17] investigated the financial performance of Hungarian and Romanian food retail companies using two clustering methods: K-Mean and K-Medoid. The paper highlights that the choice of clustering method significantly influences the assessment of financial performance, with K-Means producing a wider variety of groups and K-Medoid offering more balanced results. The research emphasizes the necessity of cluster analysis for large databases with variable quantitative data to achieve accurate results. The findings suggest that careful selection of clustering methods is crucial depending on the data and research objectives.

The study revealed that the K-Mean and K-Medoid clustering methods yield different results when evaluating the financial performance of Hungarian and Romanian food retail companies. The K-Mean method produced a greater variety of groups and a larger range of results, reflecting significant fluctuations in values. Conversely, the K-Medoid method resulted in more uniform group distributions and was less sensitive to outliers, providing a more balanced evaluation. The analysis confirmed the necessity of cluster analysis for large databases with variable quantitative data to achieve accurate results.

Huang et al. [18] presented a machine learning-based K-Means clustering method aimed at enhancing financial fraud detection in an increasingly digital financial landscape. The paper highlights the limitations of traditional rule-based detection methods, emphasizing the adaptability and precision of machine learning approaches. By clustering large volumes of transaction data, the method identifies anomalous patterns and behaviors, facilitating timely fraud detection. The research aims to improve resource allocation within financial institutions, focusing monitoring efforts on high-risk areas to mitigate fraud's impact. Overall, the study contributes to establishing a more secure transaction environment in the finance industry.

The paper demonstrates that the K-Means clustering algorithm is effective in financial fraud detection, revealing distinct clusters of fraudulent and non-fraudulent transactions. The analysis indicates that the fda model outperforms the xgbTree model in identifying fraudulent transactions, with sensitivity of 99.72 and 98.28, respectively. The clustering results show that most fraud cases are concentrated in one specific cluster, while other clusters contain minimal fraud cases. The study emphasizes the importance of cluster analysis in understanding fraud patterns and improving detection methods within financial institutions.

Deng and Mei [19] designed a clustering model V-KOSM combining SOM and K-Means clustering, which is based upon a cluster validity measure, the silhouette index. This model takes advantage of SOM where the results of SOM were applied to K-Means by avoiding one of the disadvantages of SOM (unclear clustering boundaries of nodes). As there is no consistency in the results every time the silhouette index is applied to validate the results. 100 financial statements of Chinese companies are chosen between 1999 and 2006.

47 financial ratios were selected as recognition variables. When the V-KOSM method was applied, the results ranged in accuracy from 0.79 to 0.89. In this process, only financial ratios were taken, and if non-financial ratios were taken, the accuracy could have been increased.

The experimental results indicated that the V-KSOM model achieved an average accuracy rate of 89 percent in detecting fraudulent financial statements (FFS) from the tested data. The best performance was noted with a silhouette index value of 0.2707, which corresponded to 46 correctly identified fraudulent cases out of 50. In comparison, traditional methods like hierarchical clustering and k-means clustering yielded lower accuracy rates, not exceeding 85 percent. The study emphasized the importance of using a clustering validity measure, such as the silhouette index, to enhance the model's effectiveness.

Sabau [20] surveyed clustering techniques applied in financial fraud detection over the twelve years, highlighting the increasing importance of data mining methods in combating fraud. It emphasizes the effectiveness of clustering, particularly k-means and its variants, in identifying fraudulent activities through data segmentation. The research encompasses both standalone and hybrid approaches, showcasing various applications and methodologies used in the literature. The findings underline the necessity of understanding fraud definitions and taxonomies to enhance detection and prevention strategies.

The paper surveys various clustering techniques applied in financial fraud detection over a span of twelve years, from 2000 to 2011, highlighting the increasing relevance of data mining methods in this field. It identifies k-means and its variations as the most commonly used clustering methods for outlier detection in fraudulent transactions. The research also categorizes the literature into standalone and hybrid techniques, emphasizing the importance of real datasets for quantifiable results. Additionally, it discusses the significance of understanding fraud definitions and taxonomies for effective detection and prevention strategies [9, 10].

3. RESEARCH AND DISCUSSION

In this paper, we use different clustering algorithms to compare the accuracy of the algorithms that split the data into fraudulent and non-fraudulent types. As the methodology we use is unsupervised, we assume to get less accuracy over the different algorithms used. For this purpose, we use partitioning methods (K-Means, K-Medoids (CLARA, CLARANS)), hierarchical methods (AGNES), and density-based methods (DBSCAN).

The dataset we have used is taken from GitHub and consists of 1,46,045 records, out of which 1,45,081 records are non-fraudulent, and 964 records are fraudulent. It contains fraud labels, feature variables, and related variables (e.g., fyear, gvkey, and p_aaer). The variable "misstate" is the fraud label (1 denotes fraud, and 0 denotes non-fraud). The dataset contains 28 raw accounting variables and 14 financial ratio variables. Initially, the data cleaning methods were used to preprocess the dataset. Firstly, the blank data has been removed from the dataset. Secondly the duplicate data records were also removed in the process of data cleaning.

In the related work, we have tried to generate clusters with 2 divisions (clusters) where all the objects fall into either of the 2 cluster categories fraudulent or non-fraudulent. As per the problem statement, we have used 3 different sizes of datasets, which we call low end, middle end, and high end. Depending upon execution permits, we take different counts for the above sizes that may vary over different algorithms. Low end represents the small sized dataset, middle end represents average sized dataset, and high end represents large sized dataset. Basically these 3 dataset sizes have been taken as input to check the performance of the algorithms over different sizes of the datasets. As for the results, we have taken the average accuracy of up to 5 different data instances. We have calculated the above average accuracy for 3 different sized datasets as mentioned above. Finally, we have taken overall average of the 3 sized results to get single result for each algorithm [12-15].

We have used confusion matrix to calculate accuracy among different algorithms.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where, TP = true positive, TN = true negative, FP = false positive and FN = false negative. We get the accuracy results as shown in Table 1.

Table 1. The accuracy results

Accuracy	
K Means	0.987
K Medoids	0.931
Agglomerative	0.981
CLARA	0.939
CLARANS	0.712
BIRCH	0.984

The algorithms that generate higher accuracy predict correct partitions into clusters. So, the algorithms k-Means, K-Medoids, Agglomerative, CLARA, and BIRCH performed well, which got more than 0.9 on the scale of 1 [16].

Then random index score is calculated for all the selected algorithms. The random index is calculated using the formula

$$R = \frac{a + b}{n/2}$$

where, a=the count of element pairs that belong to the same cluster; b=the count of element pairs that are assigned to different clusters; n/2=total count of element pairs in the dataset.

The results are tabulated in Table 2.

Table 2. Results of RIS

Random Index Score (RIS)	
K Means	0.974
K Medoids	0.872
Agglomerative	0.964
CLARA	0.893
CLARANS	0.653
BIRCH	0.969
DBSCAN	0.956
OPTICS	0.982

According to random index score method, higher the random index score means more accurate clusters they generate. As per the results K-means, Agglomerative, BIRCH, DBSCAN, OPTICS generates better values for random index score above 0.9 on a scale of 1 [17].

Next adjusted random index score is calculated for the selected algorithms.

Adjusted random index score is calculated using the formula:

$$ARI = \frac{R - E}{\text{Max}(R) - E}$$

where, R=The Rand index value (as defined previously), E=The expected value of the Rand index for random clusters, Max(R)=The maximum achievable value of the Rand index (always 1).

The results are shown in Table 3.

When we check these results, we find that DBSCAN, OPTICS, and CLARANS generate optimal results. Here in calculating Adjusted random index score, less the score means more the cluster accuracy the algorithms generate [18].

Table 3. Results of ARIS

Adjusted Random Index Score (ARIS)	
K Means	0.012
K Medoids	0.018
Agglomerative	0.016
CLARA	0.013
CLARANS	0.009
BIRCH	0.014
DBSCAN	-0.009
OPTICS	0.007

Next, we tested the algorithms with silhouette score. The formula for calculating silhouette score is:

$$S(i) = \frac{b(i) - a(i)}{\max\,(a(i), b(i))}$$

where, $a(i) = \frac{1}{|C(i)|-1} \sum_{C(i),i\neq j} d(i,j)$, and $b(i) = min_{i\neq j}\left(\frac{1}{|C(i)|} \sum_{j\in C(j)} d(i,j)\right)$, $C(i)$ is the cluster assigned to the i^{th} data point and $d(i,j)$ is the distance between data points i, j.

The results for silhouette scores are given in Table 4.

Table 4. Results of SS

Silhouette Score (SS)	
K Means	0.965
K Medoids	0.884
Agglomerative	0.957
CLARA	0.894
CLARANS	0.212
BIRCH	0.961
DBSCAN	-0.825
OPTICS	0.975

Table 5. Results of DBS

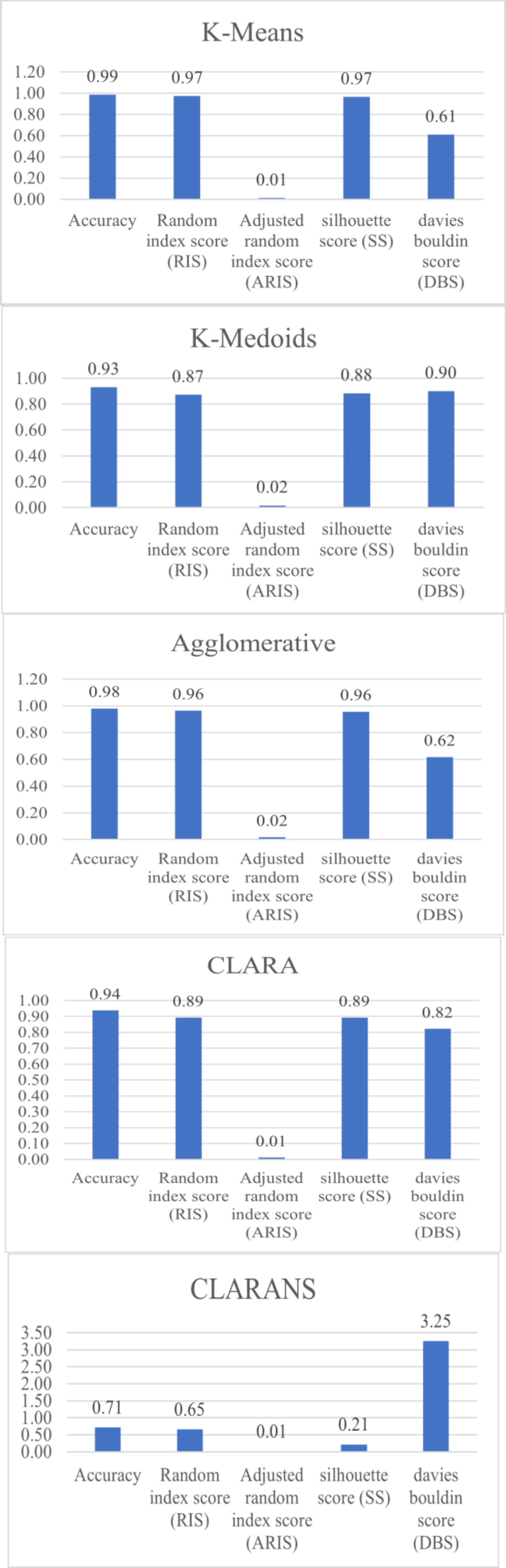
Davies Bouldin Score (DBS)	
K Means	0.609
K Medoids	0.901
Agglomerative	0.617
CLARA	0.823
CLARANS	3.251
BIRCH	0.584
DBSCAN	1.628
OPTICS	0.601

As per silhouette score method the result near to 1 represents better clustering partition and near -1 represents worst clustering partition. When we check the results, we find K-Means, Agglomerative, BIRCH, and OPTICS generate better clusters when number of clusters are taken as 2(because as per the requirement we need to take only 2 clusters fraudulent and non-fraudulent) [19].

Finally, we tested the algorithms with Davies Bouldin score method. The formula to calculate Davies Bouldin score index is

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max\left(\frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)}\right)$$

where, ΔX_k is the intra cluster distance within the cluster X_k. $\delta(X_i, X_j)$ is the intercluster distance between the clusters X_i and X_j. The results of Davies Bouldin score are given in Table 5.



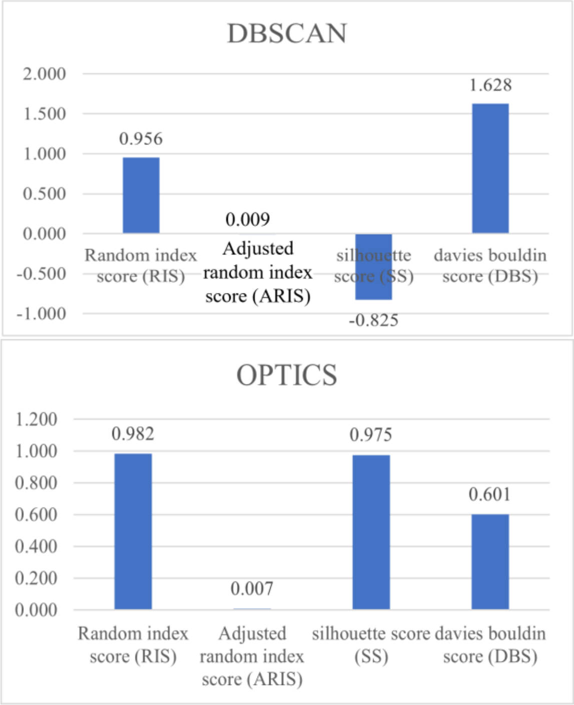


Figure 1. Clustering algorithms with their performance evaluation

As per Davies Bouldin score method less index represents good clustering, and data is well clustered. according to the results we observe that K Means, Agglomerative, BIRCH, and OPTICS generate better results over other algorithms.

We will now have an overall summary of the results. The average performance measures were recorded. The results generated by the K-Means algorithm are accuracy of 0.987, random index score of 0.974, adjusted random index score of 0.012, silhouette score of 0.965, and Davis Bouldin score of 0.609. The results of K-Medoids are accuracy 0.931, random index score 0.872, adjusted random index score 0.018, silhouette score 0.884, and Davis Bouldin score 0.901. The results of the agglomerative clustering method are accuracy 0.981, random index score 0.964, adjusted random index score 0.016, silhouette score 0.957, and Davies Bouldin score 0.617. The results of CLARA are accuracy 0.939, random index score 0.893, adjusted random index score 0.013, silhouette score 0.894, and Davis Bouldin score 0.823. The results recorded for CLARANS are accuracy 0.712, random index score 0.653, adjusted random index score 0.009, silhouette score 0.212, and David Bouldin score 3.251. The results for BIRCH are accuracy 0.984, random index score 0.969, adjusted random index score 0.014, silhouette score 0.961, and Davis Bouldin score 0.584. The results for DBSCAN are random index score 0.956, adjusted random index score -0.009, silhouette score -0.825, and Davis Bouldin score 1.628. The results for OPTICS are random index score 0.982, adjusted random index score 0.007, silhouette score 0.975, and Davis Bouldin score 0.601.

The graphical representations of the above results algorithmically are given in Figure 1. and the evaluation methodically in Figure 2.

When we consolidated all the results, we added the best performances of the algorithms in Table 6 which shows the best algorithms that perform well in different methods we compare.

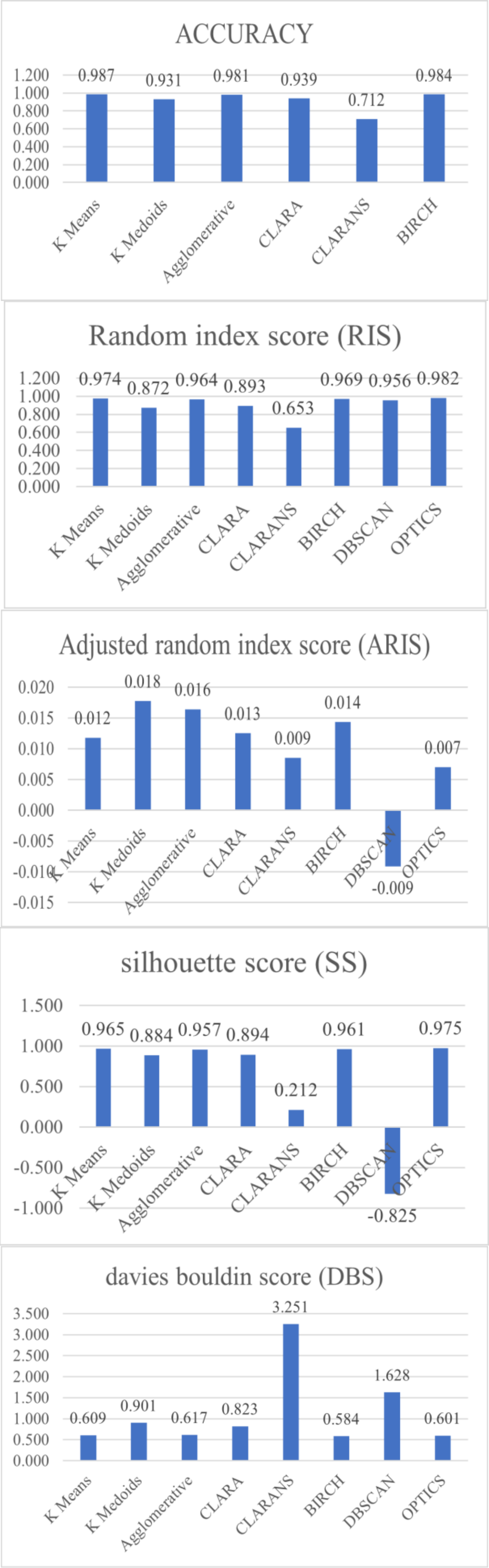


Figure 2. Performance evaluation of different evaluation methods

Table 6. The best performances of the algorithms

Algorithm	Proved Best with Methods	Total Best Performances
K-Means	accuracy, RIS, SS, DBS	4
K-Medoids	accuracy	1
Agglomerative	accuracy, RIS, SS, DBS	4
CLARA	accuracy	1
CLARANS	ARIS	1
BIRCH	accuracy, RIS, SS, DBS	4
DBSCAN	RIS, ARIS	2
OPTICS	ARIS, SS, DBS	3

Note: RIS: Random Index Score; ARIS: adjusted random index score; SS: Silhouette Score; DBS: Davies Bouldin score

So, from above table we find that K-Means, Agglomerative, BIRCH, and OPTICS algorithms generate better clusters and were proved true with at least 3 methods out of 5 used for testing. Moreover, we remove K-Means and Agglomerative algorithms from the list as they do not support large sized datasets, and we are searching for better algorithms that suits well for huge datasets. After overall comparison, we can conclude that among the clustering methods tested, BIRCH and OPTICS algorithms are found to be performing well over large datasets to detect the fraudulent statements over the given dataset.

4. CONCLUSION

In conclusion, this in-depth study of a wide range of clustering algorithms—including well-known ones like k-means, k-medoids, agglomerative clustering, CLARA, CLARANS, BIRCH, DBSCAN, and OPTICS—shows both the pros and cons of each method when it comes to finding complex patterns that are signs of financial fraud in all its forms. This study gives us a lot of useful information about the clustering techniques that work best for finding fraud in financial systems. It does this by carefully judging their performance against a wide range of criteria, such as their ability to scale, their resistance to noise, and their ability to recognize complex non-linear structures. The results of this in-depth study suggest that while traditional methods like k-means and agglomerative clustering may be better in terms of simplicity and speed, more advanced algorithms like DBSCAN and OPTICS consistently show a higher level of performance when it comes to managing and analyzing complex datasets that often have a lot of noise and different densities, which is what happens in real life. Finally, the result shows that BIRCH and OPTICS algorithms outperformed other algorithms in different accuracy reports. This large body of research ultimately makes a meaningful contribution to the ongoing efforts to improve the tools used to detect fraud in financial systems. This makes it possible to create more advanced analytical methods that can effectively adapt to the constantly changing nature of fraud activities.

REFERENCES

[1] Fahim, A. (2021). K and starting means for k-means algorithm. Journal of Computational Science, 55: 101445. https://doi.org/10.1016/j.jocs.2021.101445

[2] Patel, A., Singh, P. (2012). New approach for K-mean and K-medoids algorithm. International Journal of Computer Applications Technology and Research, 2(1), 1–5. https://doi.org/10.7753/IJCATR0201.1001

[3] Tokuda, E.K., Podruczna, A., Comin, C. H., Costa, L.d.F. (2022). Revisiting agglomerative clustering. Physica A: Statistical Mechanics and Its Applications, 585: 126433. https://doi.org/10.1016/j.physa.2021.126433

[4] Nguyen, Q., Rayward-Smith, V.J. (2011). CLAM: Clustering large applications using metaheuristics. Journal of Mathematical Modelling and Algorithms, 10(1): 57–78. https://doi.org/10.1007/s10852-010-9141-1

[5] Xu, X., Ester, M., Kriegel, H.-P., Sander, J. (1998). A distribution-based clustering algorithm for mining in large spatial databases. In Proceedings of the 14th International Conference on Data Engineering, Orlando, FL, USA , 324-331. https://doi.org/10.1109/ICDE.1998.655795

[6] Madan, S.K., Dana, K.J. (2016). Modified balanced iterative reducing and clustering using hierarchies (m-BIRCH) for visual clustering. Pattern Analysis and Applications, 19(4): 1023-1040. https://doi.org/10.1007/s10044-015-0472-4

[7] Dubey, D.S. (2023). A study on density based spatial clustering of applications with noise. ScienceOpen Posters. https://doi.org/10.14293/p2199-8442.1.sop-.pquwpp.v1

[8] Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. ACM Sigmod Record, 28(2): 49–60. https://doi.org/10.1145/304181.304187

[9] Piegorsch, W. (2020). Confusion matrix. In Wiley StatsRef: Statistics Reference Online. Wiley. https://doi.org/10.1002/9781118445112.stat08244

[10] Sundqvist, M., Chiquet, J., Rigaill, G. (2022). Adjusting the adjusted Rand Index. Computational Statistics, 38(1): 327–347. https://doi.org/10.1007/s00180-022-01230-7

[11] Xiao, J., Lu, J.,Li, X. (2017). Davies Bouldin Index based hierarchical initialization K-means. Intelligent Data Analysis, 21(6), 1327–1338. https://doi.org/10.3233/IDA-163129

[12] Shahapure, K.R., Nicholas, C. (2020). Cluster quality analysis using silhouette score. In 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, Australia, 747-748. https://doi.org/10.1109/dsaa49011.2020.00096

[13] Huang, S.Y., Tsaih, R.H., Yu, F. (2014). Topological pattern discovery and feature extraction for fraudulent financial reporting. Expert Systems with Applications, 41(9): 4360-4372. https://doi.org/10.1016/j.eswa.2014.01.014

[14] Li, T., Kou, G., Peng, Y., Yu, P.S. (2021). An integrated cluster detection, optimization, and interpretation approach for financial data. IEEE Transactions on Cybernetics, 52(12): 13848-13861. https://doi.org/10.1109/TCYB.2021.3109066

[15] Thiprungsri, S., Vasarhelyi, M.A. (2011). Cluster analysis for anomaly detection in accounting data: An audit approach. International Journal of Digital Accounting Research, 11: 69-84. https://doi.org/10.4192/1577-8517-v11_4

[16] Tatusch, M., Klassen, G., Bravidor, M., Conrad, S. (2020). Predicting erroneous financial statements using a density-based clustering approach. In 2020 the 4th International Conference on Business and Information Management, pp. 89-94. https://doi.org/10.1145/3418653.3418673

[17] Herman, E., Zsido, K.E., Fenyves, V. (2022). Cluster analysis with K-mean versus k-medoid in financial performance evaluation. Applied Sciences, 12(16): 7985. https://doi.org/10.3390/app12167985

[18] Huang, Z., Zheng, H., Li, C., Che, C. (2024). Application of machine learning-based k-means clustering for financial fraud detection. Academic Journal of Science and Technology, 10(1): 33-39.

[19] Deng, Q., Mei, G. (2009). Combining self-organizing map and k-means clustering for detecting fraudulent financial statements. In 2009 IEEE International Conference on Granular Computing, Nanchang, China, pp. 126-131. https://doi.org/10.1109/GRC.2009.5255148

[20] Sabau, A.S. (2012). Survey of clustering based financial fraud detection research. Informatica Economica, 16(1): 110-122.