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

# Intelligent Hybrid Cluster-Based Classification Algorithm For Efficient Data Analysis

**[1]Rameswara Reddy. K.V, [2]Dr. Dhyan Chandra Yadav**
[1]Assistant Professor, Department of Computer Science and Engineering,
G. Pulla Reddy Engineering College, Kurnool, Andhra Pradesh, India.
[2]Professor, Department of Computer Science and Engineering, JS University, Shikohabad,
Uttar Pradesh, India.

***Article Info***
***Page Number:685 - 696***
***Publication Issue:***
***Vol 71 No. 1 (2022)***

**Abstract:** Large amounts of data are produced by a number of industries, including financial services**,** healthcare, retail, pharmaceutical, telecom and etc. Quick processing of this large quantity of data is necessary to obtain important business insights. A large amount of data must be reacted to in real time, or nearly in real time, in order to meet the new standards. Using similarity to organize data objects into clusters, one important technique for unsupervised data processing is clustering. Multiple fields, including statistics, data mining, pattern recognition, and decision science, have examined and utilized clustering. The two primary categories of clustering algorithms are partitional and hierarchical clustering techniques. By using existing techniques, the huge amount data can't be handled without effectiveness. To generate a final clustering, a function has been used for the clustering aggregation. First, an extensive amount of basic clustering is produced by this hybrid clustering technique. Even this technique can handle large-scale datasets and extracting valuable conditions from complex data structures. Therefore, for more efficient data analysis, the intelligent hybrid cluster-based classification algorithm performs better in terms of efficiency, accuracy, and precision.

***Article History***
***Article Received: 02 January 2022***
***Revised:10 February 2022***
***Accepted: 25 March 2022***
***Publication: 15 April 2022***

**Keywords:** Clustering, Data Analysis, Hybrid Cluster, Clustering Aggregation

## I. Introduction

A collection of objects (often represented as points in a multidimensional space) can be grouped into classes of related objects through the process of clustering. The most important tool for data analysis is cluster analysis [1]. It is a collection of techniques for automatically classifying patterns according to similarity into clusters.

Patterns that belong to the same cluster intuitively look similar to each other more than patterns. It's critical to understand the differences between supervised classification and clustering, or unsupervised classification [2].

Databases can develop to be rather large, especially temporal databases and data warehouses. These databases use and usability are greatly impacted by the speed at which data can be backup. As a result, information must be arranged to allow efficient retrieving. The number of input/output operations needed to respond to a query becomes an important concern when utilizing such databases. Four common approaches are available to lower this cost: parallelism, buffering, clustering, and indexing [3]. Though any method can be used independently, clustering is clearly an essential component to the other methods.

Important steps in pattern clustering operations include pattern representation (including feature extraction and/or selection), clustering, data abstraction, output assessment, and specification of a pattern proximity suitable for the data domain. One method of exploratory discovery is cluster analysis [4]. It is useful for finding structures in data without giving an explanation or interpretation. Clustering and cluster validation are the two main components of cluster analysis. Using certain algorithms, clustering tries to divide items into groups based on specific criteria. The evaluation of the validity of clustering results by the use of clustering methodologies, algorithms, visualizations, and domain knowledge in databases is known as cluster validation, during the cluster analysis process, users can cluster and verify clusters [5]

Identifying any underlying classes in the data is another objective of clustering. In addition, clustering is a method for classifying unlabeled data into groups with little to no supervision. Because of this grouping, objects inside a class differ from one other and share properties that make them similar. Another way to set about clustering as an aspect of machine learning that addresses unsupervised learning. The learning process involves algorithms that identify patterns in datasets that have been determined from simulated or direct observation. Without knowledge of a target variable, it defined learning as attempts to categorize data observations or independent variables [6].

In the field of machine learning, clustering is a highly recognized technique known as an unsupervised learning algorithm. Sometimes the datasets cannot be processed because no class labels are available, these large amount of data is divided into smaller groups of data by the clustering approach, which makes it important. A group of data points make up each cluster, with the primary purpose of the clustering method being to categorize and assign each data point to a certain cluster [7]. Additionally, similar characteristics and/or properties should be shared by data points within the same cluster, but data points in other clusters should have extremely different features and/or properties.

Additionally, the features and/or attributes of the data points in the same cluster should be similar, but data points in other clusters should have extremely different features and/or properties.

Researchers and practitioners are becoming increasingly interested in machine learning approaches. Data clustering is a common activity in data mining and machine learning, where data is divided into groups of similar objects. Current research highlights the increasing possibility of data clustering in practical machine learning applications across different areas. Furthermore, data clustering is frequently used in data mining as a pre-processing method. However, data clustering performance is a difficult issue that researchers in several contexts have addressed in a number of disciplines [8].

Crisp and fuzzy clustering are the two general categories of data clustering. Although each data point may concurrently belong to more than one cluster throughout the clustering process, fuzzy clustering, a clustering procedure in which every data point is a part of a single cluster is referred to as crisp clustering. The most popular crisp clustering approach is the well-known Navie Bayes algorithm. Many weaknesses in this approach include instability and the high number of neighborhood searches necessary [9].

Since the 1800s, when Business intelligence (BI) first developed, data analysis has been performed for many years. However, with the development of new technology and methods for organizing, breaking down, analyzing, and extracting useful insights from that data, data analysis has undergone a significant transformation. Machine learning has been essential to those developments. Deeper and more extensive understanding can be obtained by automating the data analysis process and improving the workflow with the help of machine learning.

In order to find significant patterns and trends, data analysis requires searching and analyzing large amounts of data. By extracting useful insights from data, organizations may make better, data-informed decisions. Ultimately, data analysis can help businesses better understand their clients, develop more effective business strategies, and enhance their products and services [10].

Today, a number of tools and methods are used extensively in data analysis to support the data analysis and visualization process. These methods could include data visualization techniques, machine learning algorithms, and statistical analysis. By using these techniques, companies can obtain useful information that will aid in simplifying their processes. Later on, this data can help direct business choices, target certain consumer bases with marketing efforts, improve products services, inspire other initiatives that will enhance the company.

A subset of Artificial intelligence (AI) called machine learning using algorithms to analyze large amounts of data. Computers can effectively "learn" make predictions and decisions without explicit programming due to the development of these models and algorithms. Therefore, machine learning supports in the creation of systems that may be automatically enhanced and transformed with the addition of data or experience.

The "learned" conclusions of the computer form the basis of machine learning, as opposed to traditional programming, in which a computer scientist writes exact instructions for the computer to follow. Put another way, computers learn by observing patterns and relationships in large amounts of data that they have been trained.

In order to examine data, find patterns, and create mathematical models based on those patterns, machine learning depends on algorithms. Future events can be predicted or decided upon using the models that are produced, test hypotheses, or obtain deep insights into data that has not yet been observed or collected. Thus, in order to increase the range of data analysis and enable even stronger organizational decision-making, machine learning is proving to be essential. Many organizations are rapidly integrating machine learning into their data analysis procedures, which advances and enhances their capacity to test hypotheses and make data-driven decisions.

Three standard algorithms are used in machine learning. The process of training a model with labeled data when its final outcome or conclusion is known as supervised learning. The algorithm is able to predict new, unknown, or unlabeled data by learning from well-defined examples. The complete opposite of supervised learning is unsupervised learning. Rather of using labeled examples to train a model, the algorithm gains information through unlabeled data. Its task is to identify patterns, similarities, or clusters without any predetermined view of the final outcome. Reinforcement learning, to put it simply, instructs an agent that to interact with a novel environment and how to learn from its input. With slow improvement,

the algorithm learns from this feedback and modifies its decision-making approach to perform better over time.

Following is the arrangement of the remaining paper: In section II, The literature survey is described, section III presents the intelligent hybrid cluster-based classification algorithm for efficient data analysis; section IV explains result analysis; section V concludes the paper and references are in VI.

## Ii. Literature Survey

H. Lin and X. Xu, et.al [11] describes a Binary-digit-based data screening algorithm (BDDS) that makes use of hardware that stores data in binary form and records changes in data over time using a binary-bit recorder. The binary form's digit count is used to remove medium-distance data's effect from the current data, the decimal meaning is used to determine whether the data are valid, and the left shift of the binary form is utilized to reduce the impact of data that are very close to the current data.

S. M. Farjad and A. Arfeen, et.al [12] The suggested design examines and analyzes network traffic by utilizing statistical modeling and the clustering technique. This study provides a method for network analysis that utilizes the Packet Capture (PCAP) data format rather than NetFlow records as the main format. The clustering technique can be utilized to differentiate between malicious and benign traffic, but because of the growing application mixing and numerous dynamic circumstances, numerous questions may occur. In addition to the results of clustering, the proposed model makes use of statistical modeling.

J. Sangeetha and V. S. J. Prakash, et.al [13] The adjacency matrix is utilized by the suggested ISC algorithm to combine the clusters into a single cluster. The suggested method's effectiveness is confirmed using the existing algorithms for metrics like accuracy, memory utility, and time consumption. The compared results demonstrate that, in comparison to the current algorithms, for every measure, the suggested ISC algorithm provides the most optimal results. For big data opinion mining, an efficient method known as Inclusive similarity based clustering (ISC) is suggested. The suggested Threshold based Data Partitioning (TDP) approach is used to slice the data after it has been cleaned using the Parts of Speech (PoS) tagger.

R. Aliguliyev, A. Bagirov and R. Karimov, et.al [14] For the necessary computing time and objective function, the suggested Batch Clustering (BC) algorithm is compared to the traditional k-means clustering approach. Large data sets can be clustered using the BC algorithm in batches while retaining quality and efficiency. Numerous tests verify that the batch clustering technique, when applied to large data sets, produces better clustering than the k-means algorithm and is more efficient in terms of computing power and data storage. Two million two-dimensional data points made up the data set used in the tests.

I. Portugal, P. Alencar and D. Cowan, et.al [15] provide a framework (e.g., entry, merge, or split) for locating, processing, and analyzing interactions between spatial-temporal data clusters. They describe its structure and elements, a suggested clustering method, different techniques for measuring distance, as well as the procedure we use to calculate the cluster similarity of temporally separated clusters. The results of these processes are applied to determine the space and time relationships between clusters. The examination of these

connections reveals values that are hidden but could support novel approaches for better decision-making. Utilizing truck and human trajectories, they evaluate our framework through two case studies.

J. Hua, H. Liu, B. Zhang and S. Jin, et.al [16] introduce a novel approach called LAK(Lasso and K-means) is a computational pipeline for clustering analysis of single-cell RNA-seq data using Lasso and K-means based feature selection method, which may be used to choose candidate genes for single-cell RNA-seq data. Additionally, they made improvements to the parameter selection process based on the size of the data, binary search would automatically find the right parameters. In terms of accuracy, convenience, stability, and dependability, LAK performs better than other computational methods when applied to real datasets, simulation data, and datasets with a high number of dropout events.

M. Bendechache and M. -T. Kechadi,et.al [17] suggest a novel clustering strategy for extremely big, distributed, and heterogeneous spatial datasets. Although the method generates the number of global clusters dynamically, it is based on the K-means algorithm. Additionally, this method makes advantage of a complex aggregation step. The total procedure is optimized for time and memory allocation during the aggregation step. According to preliminary results, the suggested method scales up effectively and produces results of excellent quality. Its efficiency is significantly higher than that of two well-known clustering methods, as demonstrated by this comparison.

K. Aparna and M. K. Nair,et.al [18] An technique for bisecting K-Means based on evolutionary programming has been developed, called Canonical genetic algorithm based bisecting clustering (CGABC). The high dimensional data sets features are normalized using min-max normalization in the suggested model, and T-Test analysis comes follows. By implementing increased crossover, mutation, and a multistage reproduction process, the traditional (Genetic Algorithm) GA has been modified. The optimized cluster center information has been added in order to divide K-Means clustering, and it has proven to be an excellent method to cluster high dimensional data sets in a very accurate and effective manner.

Li P., Boubrahimi S. F. and Hamdi S. M., et.al [19] provide a novel model that uses time series data to create graphs that preserve significant relationships between various data points. Specifically, each time series data set will be treated as a node, and edges between the nodes will be added if the dynamic time warping distances of the data sets are greater than a certain threshold. Finally, the constructed graph will be subjected to the spectral clustering algorithm. This result demonstrates that the time series graph format they have proposed works better than the most advanced clustering techniques.

H. Jia and Z. Li, et.al [20] provides Spectral Ensemble Clustering with LDA-based Co-training (LSEC), a unique multi-view data clustering technique. This technique incorporates a new co-training approach into the examination of ensemble clustering. In particular, to gather refined data, they perform Linear Discriminant Analysis (LDA) on the original data using the class label that the Spectral Ensemble Clustering (SEC) obtained as the reference label, predicated on the concept that global cluster information can be obtained from the SEC clustering result. The modified data is then sent to the K-means method to obtain new basis clustering results.

C. -F. Tsai and Y. Chiang, et.al [21] creates Total-Sum-of-Squares-density-based spatial clustering of application with noise (DBSCAN), a new density-based clustering system that makes use of DBSCAN and a novel technique for applying two-phase screening, in order to improve data clustering for a number of related applications by limiting the scope of the meaningless expansion of clustering. The results of the experiments show that the novel TSS-DBSCAN scheme was faster than several popular density-based clustering techniques has very high noise filtering rate and clustering accuracy (both near to 100%). The suggested strategy might be the most effective low-time cost density-based clustering technique available presently.

P. Rathore and D. Shukla, et.al [22] The suggested clustering strategy is eventually evaluated and contrasted with the traditional k-means clustering algorithm. It was built using Java, Hadoop, and MapReduce. After the de-efficiency was eliminated, the performance that was attained demonstrated improved accuracy of cluster formation and successful consequences. As a result, the suggested work can be implemented in the big data environment to enhance clustering performance.

Du X., He Y. and Huang J. Z., et.al [23] This work presents a unique approach for handling large amounts of data clustering problems: the Random sample partition-based clustering ensemble (RSP-CE). RSP-CE is an algorithm that consists of three main parts: producing base clustering results on RSP data blocks, improving the RSP clustering results and balancing the clustering results with the Maximum mean discrepancy (MMD) criterion. Because the sample distributions of RSP data blocks are consistent across the whole dataset, applying basic clustering results on several data subsets are allows one to estimate the clustering result on the complete dataset.

D. S. B. Lalitha and S. J. Saritha, et.al [24] present the suggested incremental ensemble member selection method, the constraint propagation strategy, the automatic random clustering and subspace approach, the high dimensional facts grouping using the normalized reduction algorithm. This Incremental semi-Supervised clustering ensemble framework (ISSCE) offers a number of advantages. One of the key responsibilities and objectives of semi-supervised clustering is to organize the data devices into relevant training units (clusters) with the aim of minimizing item similarity between clusters and maximizing object similarity within clusters.

X. Gu and P. P. Angelov, et.al [25] For the processing of live data streams, they provide Autonomous Data (AD) clustering, a new autonomous data-driven clustering method. Since there are no user-defined or problem-specific assumptions or parameters required, this newly proposed algorithm is completely unsupervised entirely based on the data samples their ensemble properties. This is compared with the majority of existing clustering approaches, which have difficulty. The suggested method can also continuously update its self-defined parameters using only the current data sample, automatically developing its structure in connection with the experimentally observable streaming data, and discarding all previously processed data samples in the process.

## Iii. Framework Of Intelligent Hybrid Cluster-Based Classification Algorithm For Efficient Data Analysis

In this section, framework of Intelligent Hybrid Cluster-Based Classification Algorithm for Efficient Data Analysis is observed. A structured dataset is a group of data that has been arranged, stored, and managed for processing or analysis. Usually obtained from a single source or intended for a particular project, the data in a dataset are related in some way. The process started as the dataset has been chosen. Preparing raw data for machine learning models is known as data preparation. Making a machine-learning model started. In data science, this is the most complex and time-consuming component. To make machine learning algorithms less complex, preparation of the data is necessary. Preprocess the data and use a feature selection technique to select a few key attributes from among all the attributes. After that, principal component analysis is used to complete the feature selection.

By utilizing just relevant data and eliminating noise from the data, feature selection reduces the amount of input variable that goes into the model. Based on the type of problem that are attempting to address, it is the process of automatically selecting suitable features for the machine learning model. A common technique for reducing the dimensionality of large data sets is Principal component analysis, or PCA. It operates by dividing a large collection of variables into smaller ones that yet retain the majority of the information in the larger set. A linear dimensionality reduction method used in data preprocessing, visualization, and exploratory analysis is principal component analysis. The directions capturing the largest difference in the data are easily found by applying a linear transformation of the data onto a new coordinate system. Next, order the attributes based on expert knowledge, and determine the useful ranges for each attribute.

A machine learning model is trained using training data, which is an extremely significant dataset. Machine learning techniques are used to train prediction models with training data. A single data observation is referred to as an instance in machine learning. A data mining function that identifies target classes or categories for entries in a collection. When the learned dataset is grouped into clusters. An effective and quick method for overlapping clustering is the canopy algorithm. Because of these advantages, they adopt this approach in the multi-label classification.

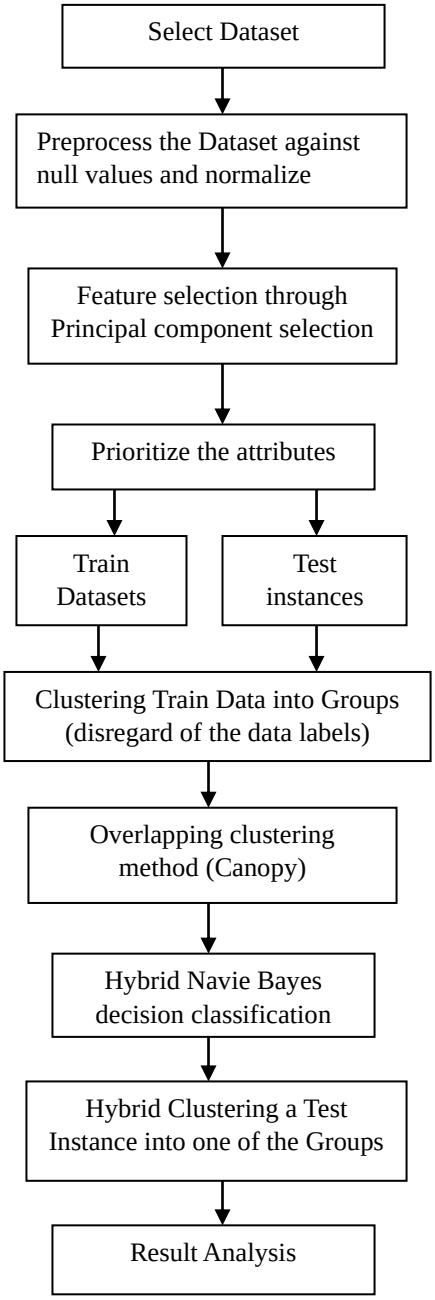


**Figure 1. Framework of Intelligent Hybrid Cluster-Based Classification Algorithm for Efficient Data Analysis**

Every instance in the proposed technique belongs to one or more classes, and multi-label classification is achieved using the overlapping clustering method. It complies with the actual situations. To evaluate the efficiency of nonoverlapping clustering through over-lapping clustering-based multi-label classification frameworks. Two well-known non-overlapping clustering algorithms are being tested against the overlapping clustering method Canopy. For decision classification hybrid navie bayes is used. A machine learning algorithm is a collection of rules or processes that an Artificial intelligence (AI) system uses to carry out tasks, most frequently to predict output values based on a certain set of input variables or to discover new patterns and insights in data. For classification tasks like text categorization, supervised machine learning algorithms like the Naïve Bayes classifier are used. They

perform classification tasks by applying probability principles. Because it makes the assumption that every input variable is independent, Naive Bayes gets the name of its creator. Although this is a significant assumption and not practical for real data, the method works extremely effectively on a wide range of challenging issues. The test instances of hybrid clustering makes into a group. Then the final result is evaluated.

## Iv. Result Analysis

In this section, result analysis of Intelligent Hybrid Cluster-Based Classification Algorithm for Efficient Data Analysis is observed.

**Table.1: Performance Comparison**

| Parameters | Batch Clustering | Hybrid Clustering |
|---|---|---|
| Accuracy | 89.6 | 94.2 |
| Efficiency | 90.1 | 99.7 |
| Precision | 85.4 | 92.3 |

In Table.1, performance comparison is observed between Batch Clustering and Hybrid Clustering interms of accuracy, efficiency and precision for Intelligent Hybrid Cluster-Based Classification Algorithm for Efficient Data Analysis.

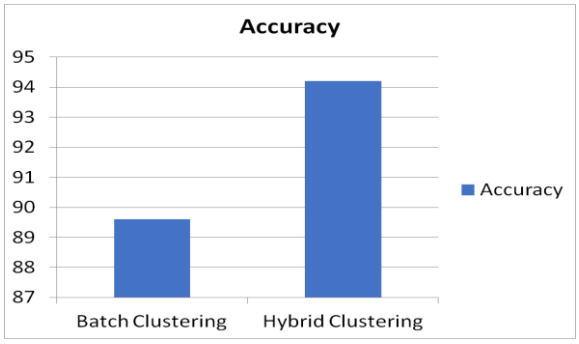


**Figure 2. Accuracy Comparison Graph**

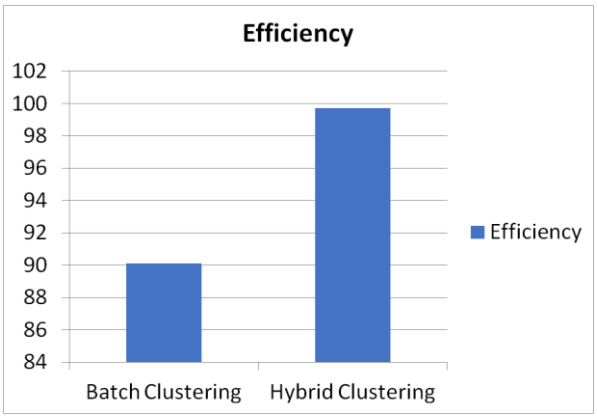


**Figure 3. Efficiency Comparison Graph**

The graphical representation between batch clustering and hybrid clustering is observed in Figure 3 for Intelligent Hybrid Cluster-Based Classification Algorithm for Efficient Data Analysis. The batch clustering shows low efficiency, when compared with hybrid clustering.
The comparison between batch clustering and hybrid clustering for precision in Figure 4 is observed for Intelligent Hybrid Cluster-Based Classification Algorithm for Efficient Data Analysis. The hybrid clustering shows high precision value.

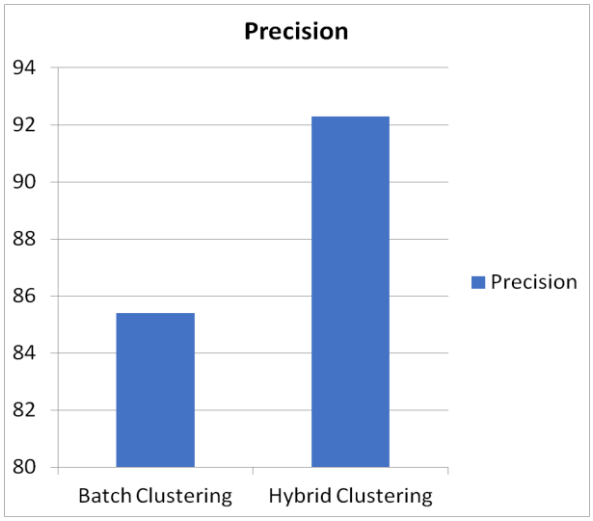


**Figure 4. Precision Comparison Graph**

## V. Conclusion

Hence, Intelligent Hybrid Cluster-Based Classification Algorithm for Efficient Data Analysis is concluded in this section. A machine learning technique called data clustering is divides a dataset into smaller sections with a higher level of intra-partition similarity and inter-partition dissimilarity. As a result, using datasets annotated by the services, this system constructs the patterns of the user behavior model. Using the classification algorithms, the framework finds utilization in the datasets. These results helped to clarify research directions and showcase the capabilities are most recent developments of available algorithms. This algorithm's primary benefit is that it offers better classification accuracy than existing systems while also helping to a reduced execution time. Even this method can handle the large data also. Hence, this method achieves better results interms of accuracy, precision and efficiency

## Vi. References

[1] A. Pandey, A. Sharma and K. K. Agrawal, "Developing efficient data mining algorithms," *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, Palladam, India, 2017, pp. 1073-1076, doi: 10.1109/ISS1.2017.8389345.

[2] Y. Wang, Y. Li, J. Sui and Y. Gao, "Data Factory: An Efficient Data Analysis Solution in the Era of Big Data," *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*, Xiamen, China, 2020, pp. 28-32, doi: 10.1109/ICBDA49040.2020.9101284.

[3] R. Shaji, "Exploratory data analysis on Reddit data: An efficient pipeline for classification of flairs," *2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM)*, Taichung, Taiwan, 2021, pp. 65-68, doi: 10.1109/BigMM52142.2021.00018.

[4] S. Aswal, N. J. Ahuja and Ritika, "Experimental analysis of traditional classification algorithms on bio medical dtatasets," *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, Dehradun, India, 2016, pp. 566-568, doi: 10.1109/NGCT.2016.7877478.

[5] D. Dhanalakshmi and A. S. Vijendran, "A novel approach in oversampling algorithm for imbalanced data sets in the context of ordinal classification," *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Chennai, India, 2016, pp. 1-5, doi: 10.1109/ICCIC.2016.7919694.

[6] Y. Jing, H. Gou, C. Fu and Q. Liu, "Study on Text Classification Algorithm Based on Non-negative Matrix Factorization," *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, China, 2017, pp. 484-487, doi: 10.1109/ISCID.2017.224.

[7] Z. Nazari and D. Kang, "A New Hierarchical Clustering Algorithm with Intersection Points," *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Gorakhpur, India, 2018, pp. 1-5, doi: 10.1109/UPCON.2018.8596795.

[8] G. Ahalya and H. M. Pandey, "Data clustering approaches survey and analysis," *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, Greater Noida, India, 2015, pp. 532-537, doi: 10.1109/ABLAZE.2015.7154919.

[9] K. M. A. Patel and P. Thakral, "The best clustering algorithms in data mining," *2016 International Conference on Communication and Signal Processing (ICCSP)*, Melmaruvathur, India, 2016, pp. 2042-2046, doi: 10.1109/ICCSP.2016.7754534.

[10] Y. Xie, A. Wulamu, Y. Wang and Z. Liu, "Implementation of time series data clustering based on SVD for stock data analysis on hadoop platform," *2014 9th IEEE Conference on Industrial Electronics and Applications*, Hangzhou, China, 2014, pp. 2007-2010, doi: 10.1109/ICIEA.2014.6931498.

[11] H. Lin and X. Xu, "BDDS: An Efficient Data Screening Algorithm Based on Binary Digit," *2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, Zhengzhou, China, 2018, pp. 492-4924, doi: 10.1109/CyberC.2018.00096.

[12] S. M. Farjad and A. Arfeen, "Cluster Analysis and Statistical Modeling: A Unified Approach for Packet Inspection," *2020 International Conference on Cyber Warfare and Security (ICCWS)*, Islamabad, Pakistan, 2020, pp. 1-7, doi: 10.1109/ICCWS48432.2020.9292396.

[13] J. Sangeetha and V. S. J. Prakash, "An Efficient Inclusive Similarity Based Clustering (ISC) Algorithm for Big Data," *2017 World Congress on Computing and Communication Technologies (WCCCT)*, Tiruchirappalli, India, 2017, pp. 84-88, doi: 10.1109/WCCCT.2016.29.

[14] R. Aliguliyev, A. Bagirov and R. Karimov, "Batch clustering algorithm for big data sets," *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, Baku, Azerbaijan, 2016, pp. 1-4, doi: 10.1109/ICAICT.2016.7991657.

[15] I. Portugal, P. Alencar and D. Cowan, "A Framework for Spatial-Temporal Trajectory Cluster Analysis Based on Dynamic Relationships," in *IEEE Access*, vol. 8, pp. 169775-169793, 2020, doi: 10.1109/ACCESS.2020.3023376.

[16] J. Hua, H. Liu, B. Zhang and S. Jin, "LAK: Lasso and K-Means Based Single-Cell RNA-Seq Data Clustering Analysis," in *IEEE Access*, vol. 8, pp. 129679-129688, 2020, doi: 10.1109/ACCESS.2020.3008681.

[17] M. Bendechache and M. -T. Kechadi, "Distributed clustering algorithm for spatial data mining," *2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)*, Fuzhou, China, 2015, pp. 60-65, doi: 10.1109/ICSDM.2015.7298026.

[18] K. Aparna and M. K. Nair, "A pragmatic approach for multidimensional data clustering," *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2017, pp. 1-6, doi: 10.1109/ICCCNT.2017.8203928.

[19] P. Li, S. F. Boubrahimi and S. M. Hamdi, "Graph-based Clustering for Time Series Data," *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 2021, pp. 4464-4467, doi: 10.1109/BigData52589.2021.9671398.

[20] H. Jia and Z. Li, "Spectral Ensemble Clustering with LDA-based Co-training for Multi-view Data Analysis," *2021 17th International Conference on Computational Intelligence and Security (CIS)*, Chengdu, China, 2021, pp. 367-371, doi: 10.1109/CIS54983.2021.00083.

[21] C. -F. Tsai and Y. Chiang, "Enhancement of data clustering using TSS-DBSCAN approach for data mining," *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, Jeju, Korea (South), 2016, pp. 535-540, doi: 10.1109/ICMLC.2016.7872944.

[22] P. Rathore and D. Shukla, "Analysis and performance improvement of K-means clustering in big data environment," *2015 International Conference on Communication Networks (ICCN)*, Gwalior, India, 2015, pp. 43-46, doi: 10.1109/ICCN.2015.9.

[23] X. Du, Y. He and J. Z. Huang, "Random Sample Partition-Based Clustering Ensemble Algorithm for Big Data," *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 2021, pp. 5885-5887, doi: 10.1109/BigData52589.2021.9671297.

[24] D. S. B. Lalitha and S. J. Saritha, "A novel data clustering through ISSCE framework," *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, Chennai, India, 2017, pp. 3595-3598, doi: 10.1109/ICECDS.2017.8390132.

[25] X. Gu and P. P. Angelov, "Autonomous data-driven clustering for live data stream," *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Budapest, Hungary, 2016, pp. 001128-001135, doi: 10.1109/SMC.2016.7844394.