

Hybrid Clustering To Big Data Analytics Related Nutrition Using Machine Learning Techniques

¹Akundi Sai Hanuman, Professor of CSE , Gokaraju Lailavathi womens engineering, Hyderabad admnglwec@gmail.com

²PM Madhuri, Student, Dept of CSE, Gokaraju Lailavathi womens engineering, Hyderabad

Abstract

Big data analytics as well as data mining are plays vital role in extracting the hidden statistics. Customary advances for investigation & extraction of hidden information from data may not exertion efficiently for big data since of its complex, very elevated volume nature. Data clustering is single of the data mining technique which exacts the useful data from the data by grouping data into clusters. In Big data as the data is complex and of very large volume, individual clustering techniques may not consider all the samples it may leads to inaccurate results. To overcome this inaccuracy this proposed method is the combination of dynamic k-means and hierarchical clustering algorithms. This proposed method can be called as hybrid method. Being hybrid method will overcome few drawbacks like static k value .In this paper proposed method is compared with existing algorithms by using some clustering metrics.

Key words: Big data, k-means, data mining, clustering

1. Introduction

Big data analytics has become trend in the market and is used to perform analytics on this big data. It is used to extract hidden patterns, unknown correlations and helps organizations in decision making. Big data is the problem and Hadoop is the solution for handling big data available as an open source framework. Clustering is one of the techniques used to extract insights from big data. Traditional clustering techniques may not work for efficient clustering in big data .consequently, there is need to plan an competent and extremely scalable clustering algorithm. This has motivated to propose a novel algorithm called hybrid clustering algorithm for big data in Hadoop ecosystem Shafeeq et.al (2012). In Big data analysis characteristics individual clustering techniques like kmeans mean and hierarchical may not consider all the samples which leads to inaccurate results Patel, D et.al (2014). K-means and hierarchical gathering techniques meet halfway because of the limitations of individual clustering algorithms. Few drawbacks of

traditional clustering algorithms are k-means clustering in this algorithm it is hard to predict the k value, wrong prediction of k value many data points may not fit into any of the clusters; several merge split decisions and iteration in hierarchical clustering etc at Paredes, G. E. (2012).

Grouping is important device for information mining & information revelation Kaur et.al (2015). The aim of bunching is to discover considerable gatherings of substances moreover to divide groups framed for a dataset Karimov, J (2015). Customary K-implies grouping functions admirably when functional to little datasets Na, S (2010). Enormous datasets should be grouped through the end objective that each and all other substance or information point in the bunch is like several elements in a similar group. Grouping issues can be applied to a few bunching disciplines. The capacity to consequently bunch comparative things empowers one to find covered up likenesses and key ideas while joining a lot of information into a couple of gatherings. This empowers clients to fathom a lot of information. Groups can be delegated homogeneous and heterogeneous bunches. In homogeneous groups, all hubs contain comparable possessions. Heterogeneous bunches are exploited in private server farms in which hubs have a variety of attributes moreover in which it could be hard to be familiar with hubs Embrechts, M. J. (2013)

Clustering techniques require the use of more exact meanings of perception and group likenesses. When gathering depends on ascribes, it is normal to utilize recognizable ideas of distance. An issue with this strategy is related with the estimation of distances between groups including at least two perceptions. Hossain, M. Z (2019) In contrast to existing regular measurable techniques, most grouping calculations doesn't depend on factual circulations of information and in this manner can be useful to apply when minimal earlier information exists on a specific issue Zaiiane (2002). Sunil Kumar. (2019) portrayed how the quantity of emphases can be diminished by parceling a dataset into covering subsets and by just emphasizing information objects inside covering zones Pande, S. R. (2012)

The remainder of this works is organized as follows. The 'History' section contains relevant surveys on the subject of Big data clustering. We provide a background on Apache Spark in 'Research Paper' The section under 'Study Design' describes the survey's research methods. The section 'Survey Methods' goes through the various Spark clustering algorithms. We provide our

analysis on clustering large data with Spark and upcoming projects in 'Discussion and Future Directions.' Lastly, in 'Findings,' bring the paper to be close.

1.1 Limitations of existing methods

The existing methods like big-data related clustering models with honeybee, genetic and PSO techniques cannot provide accurate bigdata storage. The limitations like static k, dynamic k and hadoop storage issue are cannot solve exactly. Suman (2014). The silhouette score, Calinski - Harabasz Index, and Davies - Bouldin Index cannot be improved with this method Tung-Shou Chen (2015).

2. Existing Methods

Traditional clustering algorithms like k-means and hierarchical have their own disadvantages. Data mining and big data are techniques for analyzing data and extracting confidential message.

Cluster, is a common unsupervised classification technique, is widely used in data mining, artificial intelligence, and information processing. The method entails arranging single and unique points in a group such that they are either similar to one another or different to points from other clusters. The current enormous increase of data has put traditional clustering techniques to the test. As a result, many research papers suggested new clustering techniques that take use of Big Data platforms like Apache Spark, which is intended for large data handling in a decentralized and rapid manner. Spark-based cluster study, on the other hand, is still in its infancy. Researchers examine the current Spark-based clustering techniques in terms of their support for Big Data features in this comprehensive study. In addition, we offer a new taxonomy for clustering techniques centered on Spark. To the best of our knowledge, no survey on Spark-based Big Data cluster has been performed. As a result, the goal of this study is to provide a thorough overview of past research in the area of Big Data clustering using Apache Spark from 2010 to 2020. This study also identifies new research areas in the realm of big data grouping. Because big data is complicated and large in volume, traditional methods to analysis and extraction do not function effectively. Data clustering is a common data mining method that organizes data into clusters and makes it simple to retrieve features from these regions. Traditional scheduling methods, such as k-means and hierarchical, are inefficient because the integrity of the groups they generate is harmed. As a result, an efficient and highly scalable clustering method is required. In this article, we propose hybrid clustering, a novel clustering

method that overcomes the drawbacks of current clustering algorithms. On the basis of precision, recall, F-measure, processing time, and correctness of findings, we compare the novel hybrid algorithm to current methods. The suggested hybrid clustering method is more accurate, with higher accuracy, recall, and F-calculate values, according to the research observations. Clustering analysis is the statistical mining job that attempts to make data easier to find, suggest, and organise. Clustering methods divide datasets into a number of clusters, each with its own set of attributes[7, 8]. Grouping, unlike categorization, is an iterative method in which appropriate means in a set of data are clustered into clusters[9] and thus represent various clusters, with objects in the same cluster group being very different from each other and objects in the same group or cluster being very similar to each other[10, 11]. The groups are only discovered that after classification algorithm has been completed[12]. K-means clustering and hierarchical clustering are various clustering methods that are used to manage big datasets, and each is described here.

2.1 Traditional K-Means Algorithm

K-means algorithm is used in clustering in this available are partitioned hooked on k-clusters such that there must be low inter-cluster comparison& high intra-cluster similarity T. Sajana (2016) Every cluster will have representative, from this point distance of all data points will be measured .if the distance is minimum from the centroid then these points will be considered in one cluster. Randomly chosen k-points may serve as initial centroids Yasodha, P (2015).

Drawbacks of k-means algorithm Kumar, D (2020) and Soumya, N (2020)are: it is a static algorithm and wrong estimation of k-value may affect the prediction in turn accuracy may affect .because of its sensitiveness to outliers clusters formed may not have good quality Saikumar K(2020).

k-implies combination is a technique for vector quantization, primarily commencing signal arranging, that plans to segment n perceptions hooked on k bunches in which all perception has a place among the bunch through the closest mean (group focuses or assembly centroid), filling in as a replica of the bunch Saikumar, K(2021). These effects in an apportioning of the information space into Voronoi cells. k-implies bunching limits inside group fluctuations (squared Euclidean distances), however not customary Euclidean detachments, which would subsist the more

troublesome Weber issue: the mean advances squared blunders, while just the arithmetical middle limits Euclidean distances V. Rajesh (2020). For instance, better Euclidean preparations can be exposed exploiting k-medians and k-medias Devaraju, VSN Kumar (2021).

The problem is computationally troublesome (NP-hard); notwithstanding, capable heuristic computations unite quickly to a nearby ideal Agila, D. G(2018). These are normally like the assumption boost calculation for combinations of Gaussian circulations through an iterative refinement approach exploited by mutually k-implies and Gaussian blend displaying Abinaya, S(2017). The two of them use bunch focuses to demonstrate the in sequence; be that as it may, k-implies grouping will in universal discover groups of practically identical spatial degree, while the Gaussian blend model permits bunches to have a variety of shapes Sheshasaayee, D., & Megala, R. (2017).

2.2 Hierarchical Clustering Algorithm

Hierarchical clustering forms the clusters either by combining the small clusters into large clusters or larger cluster into smaller ones. In this algorithm a tree of clusters shows the relation between the clusters are related. Hierarchical clustering may be agglomerative or divisive. And the time complexity is $O(n^2)$.

Divisive clustering initially it pools all the objects into one cluster then this initial cluster will be successive splits to obtain the divide clusters. These iterations are executed awaiting the preferred number of clusters is obtained. Its complexity is quadratic. Because of many iteration involved it takes additional time than k-means algorithm this is the main drawback of this algorithm. This procedure is by and large utilized for bunching a populace into various gatherings. A couple of normal models incorporate dividing clients, bunching comparable archives together, suggesting comparable melodies or films, and so on. There are a LOT more utilizations of solo learning. On the off chance that you run over any intriguing application, don't hesitate to share them in the remarks segment underneath!

Presently, there are different calculations that assist us with making these bunches. The most ordinarily utilized grouping calculations are K-implies and Hierarchical bunching. There are mostly II kinds of hierarchical clustering:

1. Agglomerative hierarchical clustering

2. Divisive Hierarchical clustering

2.3 Agglomerative Hierarchical Clustering

We allocate every highlight a personality bunch in this strategy. suppose there are 4 information focuses

2.4 Divisive Hierarchical Clustering

Disruptive progressive bunching works in a contrary manner. Rather than beginning with n groups (if there should be an occurrence of n perceptions), we start with a solitary bunch and allocate every one of the focuses to that group.

3. PROPOSED SYSTEM

The proposed system aims at implementing a hybrid algorithm which is a combination of Dynamic k-means and hierarchical clustering algorithms. The implemented hybrid algorithm will be run on the big datasets which gives clusters as results.

The execution of the proposed method will follow the following steps.

Step1. Generation of the dataset

Step2. Dynamic k-means will be executed for the given data set and it gives optimal value of k.

Step3. The k value obtained in step 2 is passed to agglomerative clustering for forming clusters.

Step4. The performance of the hybrid algorithm is calculated using clustering metrics.

Step5. Comparison of hybrid algorithm with existing algorithms will be performed.

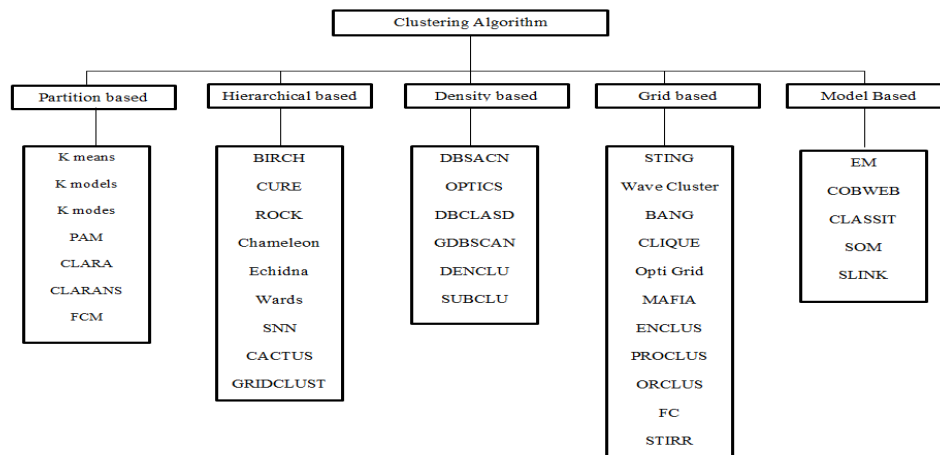


Figure 1 - Illustration of available clustering algorithms for Big Data Mining

3.1 METHODOLOGY

3.2 Flow Chart

The flow chart shown in the below figure (Figure 2) gives the flow of total process from Dataset generation to forming the clusters and finding the performance of clusters using Metrics and comparing hybrid algorithm with existing algorithm.

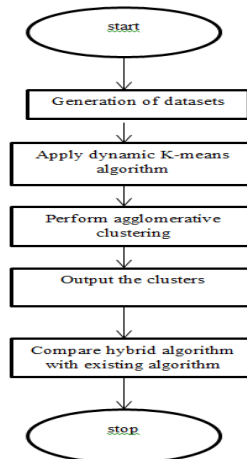


Figure 2 - Flow chart of proposed system

3.3 Dynamic k-means algorithm

Gap static method is used to get the optimal value of k . This gap statistic will compare the total intra cluster variation for various values of k . The gap statistic for a given k is defined as follows shown in eq 1

$$Gap_n(K) = E * _n \log(W_k) - \log(W_k) \text{-----} (1)$$

$E*_n$ is characterized by means of bootstrapping by producing B duplicates of the reference datasets and, by registering the normal $\log W_k$. The whole measurement estimates the deviation of the noticed W_k esteem from its normal worth under the invalid speculation. The gauge of the ideal groups will be the worth that amplifies $Gap(k)$.

The algorithm involves the following steps:

Step1. The observed data will be formed in to clusters by changing the k value from 1 to maximum. With this we can calculate the consequent W_k .

Step2. B reference data sets will be generated, cluster every of them by considering k value from 1 to maximum .

Step 3. Let $\bar{w} = \left(\frac{1}{B}\right) \sum_b \log W_{kb}^*$

Calculate the average deviation & define shown in eq 2 and 3

$$sd(k) = \left[\left(\frac{1}{B}\right) \sum_b (\log W_{*kb} - \bar{w})^2 \right]^{1/2} \text{-----} (2)$$

$$S_k = \sqrt{1} + \frac{1}{B} sd(k) \text{-----} (3)$$

4. The selected number of clusters should have smallest k such that $\text{Gap}(k) \geq \text{Gap}(k+1)$ -sk+1 Dynamic k-means algorithm is executed on the data which is read in the reduce phase and it outputs the optimal value of k which is passed to agglomerative clustering in next step.

3.3.1 Hierarchical clustering

Hierarchical clustering involves generating clusters that have a prearranged ordering commencing top to bottom. In this proposed method agglomerative clustering algorithm is used which performs the clustering in bottom-up approach. Agglomerative clustering is performed on the data also the number of clusters (k) is the output of dynamic k-means algorithm. Clusters are formed after performing agglomerative clustering.

3.3.2 Metrics for calculating the performance of clustering algorithm

The proposed method is evaluate with following metrics like Silhouette Coefficient(SC), Calinski-Harabasz index(CHI), Davies-Bouldin index(DBI).

3.3.3 Silhouette Coefficient (SC)

The silhouette plot [9] displays determine of how close every point in one cluster is to points in the neighboring clusters. In this method range of evaluates (-1, 1). This metric will be defined for each and every sample. Outcome of this metric contains two scores in that score1 if we consider it as x it will provide mean distance among from a sample to all the remaining points in same cluster. The score 2 representing it with y and it provides the mean distance between a sample and every point in the nearest next cluster.

If Considered for a single sample then $SC=y-x/\max(x,y)$. This s will be higher than the DBSCAN value.

3.3.4 Calinski-Harabasz index (CHI)

The Calinski-Harabasz index (CHI) can be referenced as Variance Ratio Criterion. The Higher score will provide better clusters. The index is the ratio of the sum of connecting-clusters dispersion and of inter cluster dispersion for all clusters. The index will have higher values for convex clusters.

3.3.5 Davies-Bouldin index (DBI)

The lower Davies-Bouldin index (DBI) specify the better parting between the clusters. This DBI index shows the average 'similarity' between clusters. The least possible score is Zero, it indicates better partition.

4. RESULTS AND DISCUSSION

The results are the screenshots of the outputs at different stages of proposed algorithm. The following shows the various stages that are present in the algorithm.

4.1 Running the algorithm in Hadoop ecosystem

Hadoop streaming jar is a utility that comes with the Hadoop distribution will have Hadoop streaming jar it is utility which allow us to create and run Map - Reduce jobs by using executable or script as the mapper or the reducer. For the implementation of the proposed method Hadoop streaming jar is used and the algorithm is written in python language. Figure 3 shows the streaming jar command used to run the python program.

```
C:\Users\SPECTRE\Downloads\hadoop-3.1.0\bin>hadoop jar C:\Users\SPECTRE\Downloads\hadoop-3.1.0\share\hadoop\tools\lib\hadoop-streaming-3.1.0.jar -input ./dataset\data.csv -
output ./hybrid_clustering_output -mapper "python E:\hadoop\hcnop.py" -reducer "python E:\hadoop\hcnred.py"
packageJobJar: [/C:/Users/SPECTRE/AppData/Local/Temp/hadoop-unjar18348713548548141475/] [] C:\Users\SPECTRE\AppData\Local\Temp\streamjob5782432461212544953.jar tmpDir=null
2020-04-08 19:19:50,087 INFO client.RWProxy: Connecting to ResourceManager at /0.0.0.0:8032
2020-04-08 19:19:50,389 INFO client.RWProxy: Connecting to ResourceManager at /0.0.0.0:8032
2020-04-08 19:19:50,620 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/SPECTRE/.staging/job_1586353545892_0001
2020-04-08 19:19:51,355 INFO mapred.FileInputFormat: Total input files to process : 1
2020-04-08 19:19:51,494 INFO mapreduce.JobSubmitter: number of splits:2
2020-04-08 19:19:51,549 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
2020-04-08 19:19:51,845 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1586353545892_0001
2020-04-08 19:19:51,848 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-04-08 19:19:52,110 INFO conf.Configuration: resource-types.xml not found
2020-04-08 19:19:52,111 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2020-04-08 19:19:52,578 INFO impl.YarnClientImpl: Submitted application application_1586353545892_0001
2020-04-08 19:19:52,784 INFO mapreduce.Job: The url to track the job: http://DESKTOP-HK080LE:8088/proxy/application_1586353545892_0001/
2020-04-08 19:19:52,812 INFO mapreduce.Job: Running job: job_1586353545892_0001
2020-04-08 19:20:03,060 INFO mapreduce.Job: Job job_1586353545892_0001 running in uber mode : false
2020-04-08 19:20:03,063 INFO mapreduce.Job: map 0% reduce 0%
2020-04-08 19:20:11,285 INFO mapreduce.Job: map 100% reduce 0%
2020-04-08 19:20:31,406 INFO mapreduce.Job: map 100% reduce 100%
2020-04-08 19:20:59,729 INFO mapreduce.Job: Job job_1586353545892_0001 completed successfully
2020-04-08 19:20:59,877 INFO mapreduce.Job: Counters: 53
```

Figure 3 running the algorithm in Hadoop ecosystem

4.2 Visual Representation of Dataset

Figure 4.2 shows the scatter plot representation of the dataset generated. The dataset is generated using make_blobs module present in sklearn and is converted into a comma-separated value (csv) file. This file is then loaded into Hadoop ecosystem through the “copyFromLocal” command. This file is read by Map phase which converts the file into a keyvalue pair and sends the data to Reduce phase. Reduce phase gets the data from Map phase and visualizes the data in the form of scatter plot.

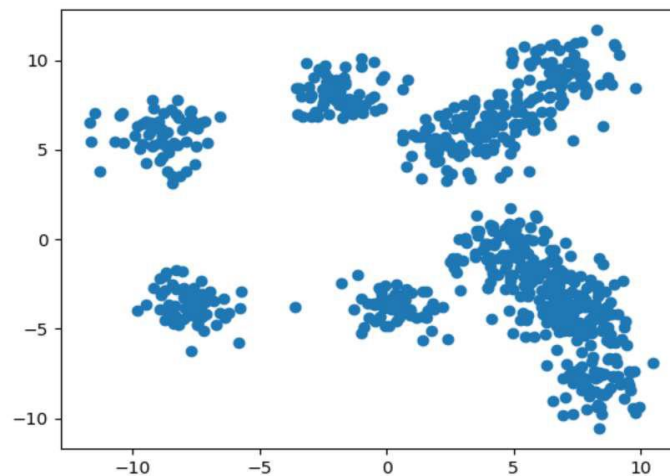


Figure 4 –Visual Representation of Dataset

4.3 Dynamic k using Gap – Statistic method

After loading the dataset, dynamic k –means is used to find the optimal value of k for a given dataset. Dynamic k –means is implemented by using gap statistic method. In this method, we choose the k value which has the maximum value of gap. Figure 4. shows the graphical Representation of gap values for various k. They used a dataset from the American Climate Server Farm (ACSF), which has the world's biggest active collection of meteorological data[25], to implement the suggested hybrid clustering method in Hadoop[24]. It provides weather files in standard ASCII format that are accessible to everyone around the world. This worldwide database combines surface hourly data from over 20,000 sites across the globe. Weather data for various years dating back to 1901 may be found in the NCDC database. Every day of the year, the temperature is documented. The input dataset for a meteorological file chosen for a certain

year looks like the image below, which displays the weather file for 1907. Figure 5 shows a short explanation of each of the 32 characteristics. There are three parts to the dataset: a control section, a required data portion, and an extra data section, all of which are detailed below. Every record begins with a 60-character corrected control signal. The control part includes data regarding the report, such as the observation date, time, and station location, among other things. Table 1 provides a short overview of each characteristic in the control section. The control section is followed by a 45-character obligatory section, which is likewise of a set length. This section includes climatic parameters such as temperature, pressure, and winds, among other things. Table 1 also includes a short explanation to each characteristic in the required section. After the required part, there is an extra data section with a variable amount of letters and no set length. It is not required, thus a given record just has to include two parts (control and obligatory). After the extra data part, a comment or element quality section may be added. In MapReduce parallel processing, the suggested method involves two main phases. The result of the mapping phase is given into the second step as an argument.

Enormous information are huge volumes, raised speed or potentially rapid data sets that include new kinds of taking care of to enhance measures, find comprehension and settle on decisions. Information catch, stockpiling, assessment, sharing, searches and representation face incredible difficulties for enormous information. Representation could be considered as "huge data front end. There's no information perception legend.

1. It is essential to imagine just brilliant data: a simple and quick view can show something off base with data very much like it assists with distinguishing energizing examples.
2. Visualization consistently shows the right decision or intercession: representation is certifiably not a substitute for basic reasoning.
3. Visualization brings confirmation: information are shown, not appearance an accurate picture of what is fundamental. Perception with different effects can be controlled.

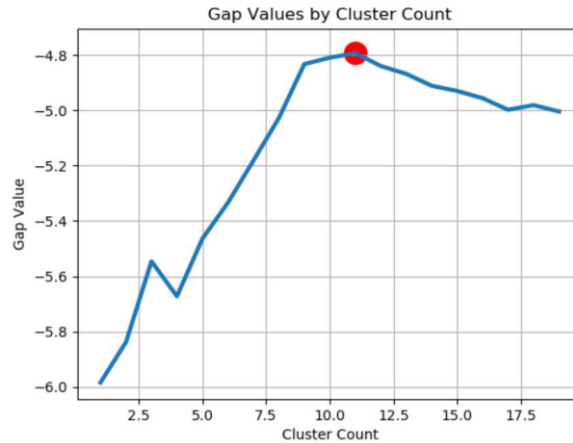


Figure 5 –Graphical representation of gap values for various k

4.4 Clusters

After getting the optimal value optimal value of k from dynamic k –means algorithm, we pass this value to Agglomerative clustering algorithm to generate the clusters. Figure 6 shows the visual representation of clusters formed after the application of hybrid algorithm.

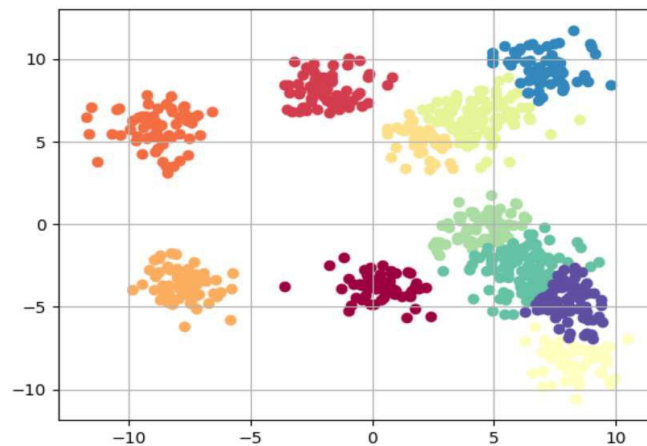


Figure 6 - Clusters

In Figure 7 the comparison of the silhouette score of hybrid algorithm and the existing algorithms i.e., k –means, hierarchical clustering is plotted. The algorithm with the highest silhouette score is more efficient compared to others.

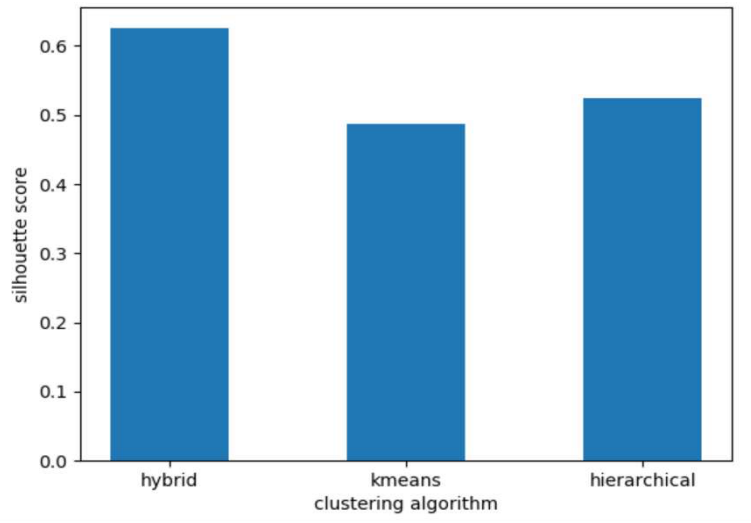


Figure 7 –Silhouette Score

Each method has disadvantages; for example, kmeans creates only a few groupings and necessitates pre-defining the number of nodes to be formed because it is dynamic in nature, whereas cluster analysis is dynamic in nature and produces more groupings than k-means, but it necessitates many iterations due to the need for many merges and split decisions. Due to these issues, we merged the two algorithms to get the benefits of each while ignoring the drawbacks. We discover the greatest number of clusters from a file using the resultant hybrid method, and the cluster produced are of extremely high quality, resulting in its most including among. The suggested hybrid technique results in a more effective cluster analysis with improved accuracy, recall, and Fmeasure. Because the computed highest temperature value equals the real high temperature value, the result generated by the efficient hybrid clustering method is the most reliable. The hybrid method generates the most clusters and incorporates all data points in either of these groups.

4.5 Calinski–Harabasz Index

The Figure 4.6 shows the comparison of the Calinski–Harabasz Index of hybrid algorithm and the existing algorithms i.e., k –means, hierarchical clustering. The algorithm with the highest Calinski–Harabasz Index is more efficient compared to others.

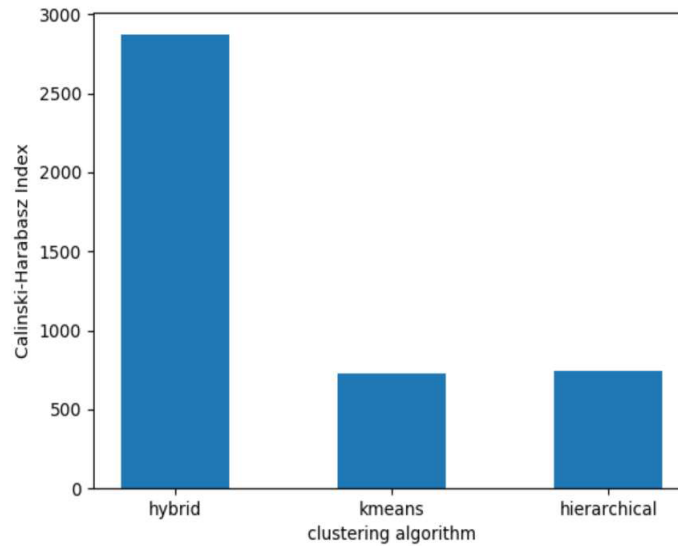


Figure 8 –Calinski–Harabasz Index

The Figure 8 and 9 shows the comparison of the Davies–Bouldin Index of hybrid algorithm and the existing algorithms i.e., k –means, hierarchical clustering. The algorithm with the least Davies Bouldin Index is more efficient compared to others.

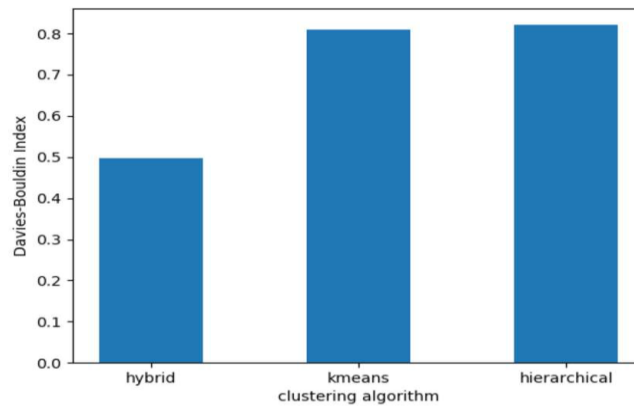


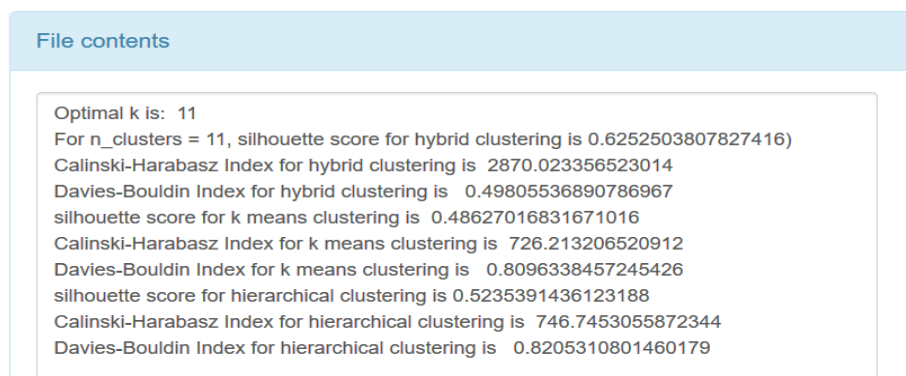
Figure 9–Davies –Bouldin Index

Stage of the mapper the meteorological file for a specific year is sent into the Mapper phase, as seen in Fig. 3 for the year 1907. Every record has two fields: the measurement date and the ambient temperature. The quality code is also checked to ensure that its value isn't missing. The Mapper splits the extracted data into key-value pairs, since the MapReduce concept is built on a key-value situation. The date (as a key) and temperatures (as a value) are sent to the reducer phase, with the temperature being reduced by ten. The value is in IntWritable form, while the key is in Text form. The such are for this data are (Monitoring date, Wind direction (Celsius));

some examples are (19070101, 13.9), (19070102, 11.7), and (19070103, 11.9). (19070103, 12.3). Each keyvalue pair is distinct. Figure 4 shows an output data with the two fields retrieved in the Analyzer phase date and observed surface temperature on that day. Decrease the number of stages. The reduction phase takes the Mapping phase's result as its argument. It accepts key-value pairs in two formats: Text and IntWritable. That is, all of the temp readings were taken on a given day, at a particular time, and under precise circumstances. The temperature may be monitored quite often on a given day, but only up to three times in a single day, thus each key value is unique. The speed of the Enhancer varies depending on the method used to identify similar and determine the maximum temperature from those groups.

4.6 Final Output

The final output consists of the optimal no. of clusters and the above mentioned clustering metrics i.e., Silhouette score(SC), Calinski–Harabasz Index(CHI), Davies –Bouldin Index(DBI) for hybrid clustering algorithm, k –means and hierarchical clustering algorithms. We can view them inHadoop environment directly using “cat” command or can view them using“http://localhost:9870/”. “http://localhost:9870/”contains the details of the Hadoop Environment such as the no. of nodes used, amount of space used, details about node manager, resource manager etc. We can also the view the various files present in the Hadoop environment. Our algorithm stores the final outputs in a file in Hadoop ecosystem. Figure 8 shows the content the file which contains the final outputs.



```

File contents

Optimal k is: 11
For n_clusters = 11, silhouette score for hybrid clustering is 0.6252503807827416)
Calinski-Harabasz Index for hybrid clustering is 2870.023356523014
Davies-Bouldin Index for hybrid clustering is 0.49805536890786967
silhouette score for k means clustering is 0.48627016831671016
Calinski-Harabasz Index for k means clustering is 726.213206520912
Davies-Bouldin Index for k means clustering is 0.8096338457245426
silhouette score for hierarchical clustering is 0.5235391436123188
Calinski-Harabasz Index for hierarchical clustering is 746.7453055872344
Davies-Bouldin Index for hierarchical clustering is 0.8205310801460179

```

Figure 10 –Final Output

The percentage of pairs of data points properly put in same clusters is used to calculate accuracy. It is proportional to the quality of clusters produced and reliability; the lower the sensitivity, the lower the quantity of groups defined; the greater the specificity, the more precise

the algorithms is, and the better the grade of groups formed. Precision is defined as the number of clusters calculated using a certain method. The dataset's real groups the variation from the true value is used to calculate it. Precision is calculated by dividing the number of clusters produced by a given method by the total number of clusters that may be formed in the dataset. We may also refer to it as the application's proportion of meaningful clusters. The hybrid method has the greatest accuracy, whereas the k-means approach has the least precision, as seen in Figure 5. The time value of money is evaluated between the three clustering methods. The execution times of the three methods are compared in Figure 8. The k-means process utilizes the least amount of time to create clusters, whereas the hybrid approach takes the best time.

5. CONCLUSION

The proposed hybrid algorithm will overcome the drawbacks of existing algorithms and produce efficient results. This approach is an efficient way of producing the clusters. This hybrid algorithm is compared with existing algorithm using clustering metrics such as silhouette score, Calinski - Harabasz Index, Davies - Bouldin Index and proved that proposed algorithm is efficient than existing algorithms. Due to our system configurations, we could not test for very large datasets but it can be extended in future.

REFERENCES

- Shafeeq, Ahamed, and K. S. Hareesha. "Dynamic clustering of data with modified k-means algorithm." Proceedings of the 2012 conference on information and computer networks. 2012.
- Patel, D., Modi, R., & Sarvakar, K. (2014). A comparative study of clustering data mining: techniques and research challenges. *International Journal of Latest Technology in Engineering, Management & Applied Science*, 3(9), 67-70.
- Paredes, G. E., & Vargas, L. S. (2012, June). Circle-Clustering: A new heuristic partitioning method for the clustering problem. In *The 2012 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.
- Kaur, Jaskaranjit, and Harpreet Singh. "Performance evaluation of a novel hybrid clustering algorithm using birch and K-means." 2015 Annual IEEE India Conference (INDICON). IEEE, 2015.
- Karimov, J., & Ozbayoglu, M. (2015, October). High quality clustering of big data and solving empty-clustering problem with an evolutionary hybrid algorithm. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 1473-1478). IEEE.

Na, S., Xumin, L., & Yong, G. (2010, April). Research on k-means clustering algorithm: An improved k-means clustering algorithm. In 2010 Third International Symposium on intelligent information technology and security informatics (pp. 63-67). Ieee.

Embrechts, M. J., Gatti, C. J., Linton, J., & Roysam, B. (2013). Hierarchical clustering for large data sets. In *Advances in Intelligent Signal Processing and Data Mining* (pp. 197-233). Springer, Berlin, Heidelberg.

Hossain, M. Z., Akhtar, M. N., Ahmad, R. B., & Rahman, M. (2019). A dynamic K-means clustering for data mining. *Indonesian Journal of Electrical Engineering and Computer Science*, 13(2), 521-526.

Zaïane, Osmar R., et al. "On data clustering analysis: Scalability, constraints, and validation." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Berlin, Heidelberg, 2002.
Sunil Kumar and Maninder Singh "A Novel Clustering Technique for Efficient Clustering of Big Data in Hadoop Ecosystem", *Big Data Mining and Analytics*, ISSN 2096-0654/08 pp240–247, Volume 2, Number 4, August 2019.

Pande, S. R., Sambare, S. S., & Thakre, V. M. (2012). Data clustering using data mining techniques. *International Journal of advanced research in computer and communication engineering*, 1(8), 494-9.

Suman and Mrs.PoojaMittal"Comparison and Analysis of Various Clustering Methods in Data Mining on Education data set Using the weak tool" of *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)* Volume 3, Issue 2, March –April 2014.

Tung-Shou Chen, Tzu-Hsin Tsai, Yi-Tzu Chen, Chin-Chiang Lin, "A combines k-means and hierarchical clustering method for improving the clustering efficiency of microarray", 2005 *International Symposium on Intelligent Signal Processing and Communication Systems*.

T. Sajana, C. M. Sheela Rani and K. V. Narayana "A Survey on Clustering Techniques for Big Data Mining", *Indian Journal of Science and Technology*, Vol 9(3), January 2016.

Yasodha, P., & Ananthanarayanan, N. R. (2015). Analysing big data to build knowledge based system for early detection of ovarian cancer. *Indian Journal of Science and Technology*, 8(14), 1.

Kumar, D. S., Kumar, C. S., Ragamayi, S., Kumar, P. S., Saikumar, K., & Ahammad, S. H. (2020). A test architecture design for SoCs using atam method. *International Journal of Electrical and Computer Engineering*, 10(1), 719.

Soumya, N., Kumar, K. S., Rao, K. R., Rooban, S., Kumar, P. S., & Kumar, G. N. S. 4-Bit Multiplier Design using CMOS Gates in Electric VLSI. *International Journal of Recent Technology and Engineering (IJRTE)* ISSN, 2277-3878.

Saikumar K, Rajesh V, Hasane Ahammad S K, Sai Krishna M, Sai Pranitha G, Ajay Kumar Reddy R, Cab for Heart Diagnosis with RFO Artificial Intelligence Algorithm , International Journal of Research in Pharmaceutical Sciences: Vol. 11 No. 1 (2020)

V Saikumar, K., Rajesh "A novel implementation heart diagnosis system based on random forest machine learning technique "International Journal of Pharmaceutical Research 12, 3904–3916

Saikumar, K., and V. Rajesh. "DIAGNOSIS OF CORONARY BLOCKAGE OF ARTERY USING MRI/CTA IMAGES THROUGH ADAPTIVERANDOM FOREST OPTIMIZATION." Journal of Critical Reviews 7.14 (2020): 591-600.

Devaraju, VSN Kumar, and Sirasani Srinivasa Rao. "A Real and Accurate Vegetable Seeds Classification Using Image Analysis and Fuzzy Technique." Turkish Journal of Physiotherapy and Rehabilitation 32: 2.

Agila, D. G., & Arumugam, D. (2018). A Study on effectiveness of promotional strategies at Prozone mall with reference to visual merchandising. International Journal of Innovations in Scientific and Engineering Research, 5(6), 47-56.

Abinaya, S., & Arulkumaran, G. (2017). Detecting black hole attack using fuzzy trust approach in MANET. Int. J. Innov. Sci. Eng. Res., 4(3), 102-108.

Sheshaaayee, D., & Megala, R. (2017). A Conceptual Framework For Resource Utilization In Cloud Using Map Reduce Scheduler. International Journal of Innovations in Scientific and Engineering Research (IJISER), 4(6), 188-190.