AMSE JOURNALS-AMSE IIETA publication-2017-Series: Advances B; Vol. 60; N°3; pp 525⁻538
Submitted July 25, 2017; Revised Dec. 26, 2017; Accepted Jan. 15, 2018

A Novel Algorithm for Clustering and Feature Selection of High Dimensional Datasets

*Thulasi Bikku, **Alapati Padma Priya

*Member, IAENG, Department of CSE (thulasi.jntua@gmail.com)
**Vignan's Nirula Institute of Technology and Science for Women,
Palakaluru Road, Guntur, Andhra Pradesh 522005, India (padmapriyaalapati01@gmail.com)

Abstract

The exponential development of perplexing, heterogeneous, dynamic, and unbounded information, produced by an group of fields including health, genomics, material science, climatology, and interpersonal organizations posture noteworthy difficulties in information preparing and desired speed-execution. The responsibility of collection and arrangement of objects such that items in a similar group are more comparable to each other than to those in different groups (clusters). Exploratory information technique is clustering, which arranges the information of dataset into a few groups. There are many grouping methods are accessible. Various types of calculations are best utilized for various types of information. K-means is mostly utilized for clustering analysis algorithm. Big data analytics includes numerous imperative data mining undertakings including clustering, which arranges the information into important clusters in view of the likeness or uniqueness among objects. Experiments are performed on a benchmark dataset to assess the attainability and effectiveness of our calculation. Immense measure i.e. Gigabytes, Terabytes) of information processing and analysis is done using the big data environment. For Cluster analysis technique, mainly the K-mean clustering algorithm is executed through the Hadoop and MapReduce to analyse high dimensional datasets. In big data analytics, the clustering is done when the unlabelled information is handled and used to group clusters of the information. Also when it is examined by the conventional k-means algorithm does not works well with the Hadoop framework and MapReduce programming in this manner it is mandatory to change the algorithm so as to improve the performance on the data analysing techniques. In this manner another clustering algorithm with improvement on conventional k-

means clustering algorithm is proposed and executed. This approach initially upgrades the quality of the data by evacuating the anomaly focuses in datasets and afterward the bi-part technique is utilized to play out the grouping. The proposed algorithm for clustering method executed utilizing the Hadoop framework and MapReduce programming at long last the execution of the proposed algorithm of grouping approach is assessed and contrasted and the conventional k-means clustering technique. The acquired execution demonstrates the compelling outcomes and improved accuracy of group construction with the evacuation of the de-effectiveness. In this way the proposed work can be applied for big data environment with enhancing the execution of grouping

Key words

K-means, Classification, Clustering, MapReduce, Big data, Accuracy, Hadoop.

1. Introduction

Nowadays, data acquisition methods and information stockpiling media have advanced quickly. This has brought about an exponential development of mind boggling, heterogeneous, dynamic, and unbounded information being created by an assortment of fields including human health, genomics, material science, climatology, social networking and interpersonal organizations. For example, in genomics, the measure of structured and unstructured information created multiplied like clockwork inside the most recent decade, now delivering a few peta-bytes of information consistently [1]. Likewise, the volume of information as of now delivered by NASA Earth science mission is around 12 peta-bytes, and is required to develop by a request of extent inside the following five years. Breaking down and preparing such a colossal measure of information posture genuine difficulties to the information mining group. Big data analytics frequently includes numerous critical information mining assignments, for example, classification, clustering, prediction, regression, and association rule mining [2]. From these, we are concentrating on the most broadly utilized classification and clustering. The vast majority of the present information mining assignments, including grouping and order, for big data analysis are ending up more unpredictable, requiring more handling power than any other time in recent memory. Likewise, by and large, the information should be prepared continuously to yield the genuine advantage. These imperatives significantly affect the speed-execution and accuracy of the data mining applications. Existing methods for huge information investigation are normally

processor-based (programming) plans. These processor based methods are incapable of analyzing and processing huge measures of information, proficiently and viably [3].

Characterization and clustering (grouping) is a conventional issue in text mining. In text mining two sorts of approaches are used, one is supervised and the other is unsupervised [4]. In unsupervised technique, unsupervised algorithms are used to mine and analyse the data, which is known as cluster analysis. Lately various commitments on document classification approaches are identified. In hierarchical strategy of document clustering, grouping is as often as possible utilized, which give much precise outcomes however this technique is lack in accuracy [5]. Then again, the K-means clustering is separately better execution method. Be that as it may, the precision and the execution of the k-means technique are fluctuating as indicated by the information and with various tests. In the demonstrated work, the k-means method is changed for discovering or perceiving the enhanced and steady execution of document grouping for enormous information perception [6]. The enhanced K-mean algorithm defeats the issues display in the enhanced k-means calculation by pre-handling the data of the datasets of real-time. The enhanced k-means algorithm is a single pass technique. Various types of calculations are best utilized for various types of information. Traditional K-means method is the essential grouping procedure of clustering and order of dataset. K-means is denoting a grammatical rule that can be applied repeatedly to cluster data into k groups. It gives best outcome and is effectively implementable [7]. Be that as it may, when it is connected on high dimensional information, the unpredictability ends up noticeably infeasible. Thus the enhanced k-means clustering algorithm is preferred to put into effect for the big data analysis effectively.

Rest of the paper describes: section II explains about Algorithm for Traditional K-Means, Section III give a detailed account of the related work about k-means algorithm, section IV portrays the proposed algorithm i.e. K-means for High Dimensional Data (KHDDD) description. In section V experimental results have been shown and a comparison has been made between the traditional k-means, improved k-means+ and proposed algorithm for k-means (KHDDD). Finally section VI tells about the conclusion and future work.

2. Algorithm for Traditional K-Means

K-means algorithm is a kind of unsupervised learning, which is used during unlabeled data (i.e., data without groups). The goal of this estimation is to find bunches in the data, with the amount of groups named by the variable K [8]. The count works iteratively to apportion each datum point to one of K groups in similitude of the features that are given. The aftereffects of the K-means grouping calculation are:

Firstly, it arbitrarily chooses k objects from the entire data which introductory cluster centers. Each outstanding object is doled out to the group to which it is the most comparative, in view of the separation between the object and the cluster center. The new mean for each group is then ascertained. This procedure repeats until the point that the criterion function unites. Every centroid of a group is a collection of feature values which characterize the subsequent groups [9]. Looking at the centroid include feature weights can be utilized to subjectively translate what sort of cluster each group speaks about:

1. Partitioned clustering approach
2. Each group is related with a centroid (mean position)
3. Each point of data is relegated to the group with the closest centroid
4. Number of groups, K must be specified

Points of interest of K-means algorithm are specified: Initial centroids are often selected arbitrarily, in view of the centroids the Clusters delivered change starting with one run then onto the next. The centroid is (normally) the mean of the focuses in the group. 'Closeness' is measured by Euclidean Distance(ED), Cosine similarity, Correlation, and so forth [10]. The vast majority of the convergence occurs in the initial couple of cycles, the ceasing condition is changed to 'Until the point when generally few focuses change clusters'.

K-MEANS using MapReduce:

1. Initially select k means $\mu_1 \ldots \mu_k$ uniformly at random from the set X.

// where μ is means of all datapoints in the group

2. Apply the MapReduce given by k-meansMap and k-meansReduce to X.
3. Compute the new means $\mu_1 \ldots \mu_k$ from the results of the MapReduce.
4. Broadcast the new means to each machine on the cluster.
5. Repeat steps 2 through 4 until the means have converged.

The number of iterations required for convergence can theoretically be quite large, but in practice it is typically a few dozen even for large datasets. The k-means algorithm has a small amount of issues, which influences its execution, these are: introduction of cluster and initialization of centers, Responsive to anomalies i.e. outliers, Fixing size of the count k, finishing groups may stay discharge i.e. no information point would be relegated to it and when dataset is substantial the time complexity nature is too high, since it relies upon both, the quantity of items and the quantity of groups.

3. Related Work

K-means strategy has a couple of issues as analyzed already. Heaps of work in the research field has been done to address the issues, and a couple of upgrades have been proposed to enhance its opportunity time complexity of the k-means algorithm [11]. Vance Faber has anticipated a way of dealing with a situation or problem, called "continuous k-means algorithm", with the help of random sampling system in which the data point initialization and data point re-assignment methods have been applied [12]. From the accessible populace of dataset the data center of the cluster focuses are chosen as random specimens in a way that the dense region in the original dataset are present, than it ought to likewise be reflected in the chosen random specimen. With the assistance of random specimen testing the information point's re-task step is likewise refined. In each progression for each group just a random specimen of its information focuses are chosen in the event that they should be reassigned. This approach brings about better outcomes and merges in the wake of checking just a little piece of the expansive information dataset. Malay K. Pakhira has done an enhanced algorithm to keep away from exhaust groups in the traditional k-means algorithm [13]. To alleviate the impact of purge groups, aside from including every one of the individuals from those groups, likewise incorporate the current group intends to ascertain new group mean. Along these lines if in any progression the group centre points are equivalent in the following stage, this incorporation of present group mean in new centre calculation will bring about various centers for the clusters. It defers the joining condition by including additional iterations. Dar-Ren, et al. exhibited another approach of taking care of expansive dataset for grouping by applying a matrix based system [14]. The region encompassed by the greater part of the data points in input dataset is separated into some equivalent size of matrices. Just nonempty matrices are decided for thought i.e. which has no less than one data point in its zone. Agents are browsed these lattices, and thought to be the dataset for the k-means algorithm. The clustering technique is keep running for these agents, considering as data points. Groups are produced from this new dataset. Every one of the information focuses contained in one lattice is spoken to by its delegate, and move along. The quantity of networks has been characterized exactly for this approach. Shehroz S. Khan suggested a method; it goes for calculating starting group center for each feature in the dataset, as opposed to for centers [15]. Along these lines outliers are not get picked and starting centers of the cluster comes about nearer to precise group centers. To accomplish this, for a collection of related sets of information that is composed of separate elements is attempted to be typical disseminated. The typical curve represented by this feature for all points of data is drawn and separated into k equal parts. The yield brings about k vectors for that features, these vectors fills in as initial centers of clusters for its particular feature in primary

k-means algorithm. These points which are considered as outliers don't get selected as initial group center points. M. P. Sebastian have proposed an alternate center for cluster initialization plot so as select center points from denser zone and to maintain a strategic distance from outliers [16]. In this technique, iteratively k sets are created, each containing storerooms from (3/4) of (n/k) items. Each set can have break even with number of elements and no items stays finally. Current methods for these subsets are ascertained which fill in as seed for instatement step of the method. Jieming Wu and Wenhu Yu have presented a preprocessing of input information to diminish the quantity of cycles in the clustering technique [17]. The thought is to first sift through those k data centers which are most disengaged information points. While choosing introductory data centers, if these confined centers get chose, the general calculation would be expanded, can misshape the state of groups. Sifting through k most far off the information items will guarantee that the data points chosen as starting cluster centers. Performing these per-handling steps, the underlying group focuses are nearer to the final centers of the cluster. Joab O. Lima have utilized multi-leveled clustering algorithm for the instatement step [18]. Standard k-means algorithm result is influenced by the decision of initial cluster center points. In the standard algorithm the underlying group center point are picked as random data point, which gives diverse outcome when run. The algorithm runs a few times by applying arbitrary initial center points, therefore gathered final group center points are reduced to k numbers by the agglomerative hierarchical mean approach. These outcomes are closer to final cluster center points. Topchy, Alexander, Anil K. Jain, and William Punch have explained a few changes to the conventional k-means algorithm in data point reassignment steps. They have utilized the way that in two progressive strides there are many points which stay unaltered and are not re-allotted. When appointing a point to a group its distance from its group is processed and stored. In next iteration if the distance from the new center point of the group for that data point is observed to be less, the point task stay unaltered, whole process for task isn't run, which brings about decrease of calculation. Be that as it may, the arrangement is appropriate just for two-dimensional information [19].

4. Proposed Algorithm (KHDD)

Data mining is a strategy of programming application, which is utilized to analyse the extremely large amount of information i.e. structured, unstructured and semi-structured data. To analysis of the information utilizing two distinct behaviors are considered that is relies upon the information; they are supervised learning and unsupervised learning. The information holds a number of connected items of features and their resultant class labels then this information can be analysed utilizing the supervised techniques, for example, decision tree, automation, neural

network system and artificial intelligence and if the information isn't sorted out through the predefined patterns then that is utilized with the unsupervised learning procedure. The supervised learning bolsters the classification methods to execute and the unsupervised learning underpins the clustering algorithms [20]. After examination of traditional K-means algorithm some of insufficiencies are identified, for example, fluctuations in precision, high error rate and others. Keeping in mind the final goal is to determine the issues in conventional K-means algorithm, so proposed another algorithm. The proposed method initially pre-processes the information and recognizes the outliers from a collection of related sets of information that is composed of separate elements. After that the information is utilized with an arrangement of procedures and their results are assessed utilizing the accurate validation and verification methods. Along these lines in this introduced work another improved approach for improving the execution of K-means algorithm is proposed. The clustering method guarantees to reduce the resource utilization and improve the group leader selection process. By which the issues of the fluctuating precision and higher error rate is lessened essentially moreover the given proposal decreases the time utilization of clustering procedure by performing less the repetition of a process with assessment of information [21].

The improved k-means algorithm (KHDD)

Every datum point contains M estimations, that is, a data point can be addressed as an ordered set of data constituting a record with M variables (V_{a1}, V_{a2}… … V_{am}). Select p estimations randomly out of M estimations of the data centers. Sort out them in lessening solicitation of need concurring the required convergence of examination, as (d_1, d_2, d_3… d_p) .Assume the principle estimation as d_1, is the basic estimation and rest of the estimations is known as optional estimations. The size of group is k, which is selected prior.

Step 1: For each dimension, the variant is computed as,$Hi = maxi - mini/k$, where maxi is the maximum value of i^{th} dimension and mini is minimum value of i^{th} dimension.

Step 2: Initially the clusters combine together using following state of affairs. For any data point if min1+J*H1<=value1<min1+(j+1)*H1, then the data point of the group belongs to cluster j.

Step 3: For each and every cluster the centroid data point is computed. Centroid is a mean of all cluster points.

Centroid= (cl_{p1}+cl_{p2}+…..Cl_{pn})/n

Step 4: The computation of secondary dimension, 2≤j≤m, repeat the following

1. To detect the outliers of each and every cluster based on the following clauses.
2. For every data point If$|valueij - valuecj| > Hj$

3. Then the data point is outlier based on dimension j. Here, valueij is value of the ith data point's jth dimension, and valuecj is value of the centroid's jth dimension.

4. Compute distance from each centroid ith to jth dimension of outlier data point.

5. As distance $=\sum_{l=1}^{l=j}|\text{valueij} - \text{valuecj}|$.

6. Choose the cluster of the data point based on least distance from all data points.

5. Experimental Results

This segment gives the outcomes, accuracy and the proposed algorithm's achievement, with reference of K-mean and k-means+ clustering algorithm. The execution of the method is assessed as far as precision, error rate. To appraise outcomes about there are 4 diverse datasets are utilized and the significant execution of the framework is given underneath.

Entropy measure is processed inside the cluster to assess the document's group consistency. As the entropy and average separation index are high, at that point the quality illness designs have high consistency and least error inside the candidate patterns [22]. Average separation index that measures the greatness of gaps between any two clusters in a segment, by anticipating the information in a couple of groups into a one-dimensional space in which they have the most extreme partition. Table 1 shows the comparison between different algorithms and the proposed KHDD model, which had given good results.

Table 1. Comparing different instances based on Average Separation Index

	Average separation index (with 5 instances)			**Average separation index (with 10 instances)**		
Data set	**K-Means**	**K-Means+**	**Proposed method (KHDD)**	**K-Means**	**K-Means+**	**Proposed method (KHDD)**
Abalone	0.85722	0.82425	0.785	0.80546	0.78049	0.76145
Brest cancer	0.83101	0.79905	0.761	0.78084	0.75662	0.73817
Forest fire	0.81245	0.7812	0.744	0.76339	0.73972	0.72168
Iris	0.86159	0.82845	0.789	0.80957	0.78446	0.76533

The Separation Index is measured for various calculations with various datasets are thought about. Hence the results prove that the proposed algorithm is efficient than the conventional algorithms such as k-means and k-means+ [23]. Figure 1 shows the graphical portrayal of Average Separation Index of the different models on different datasets compared with our KHDD model.

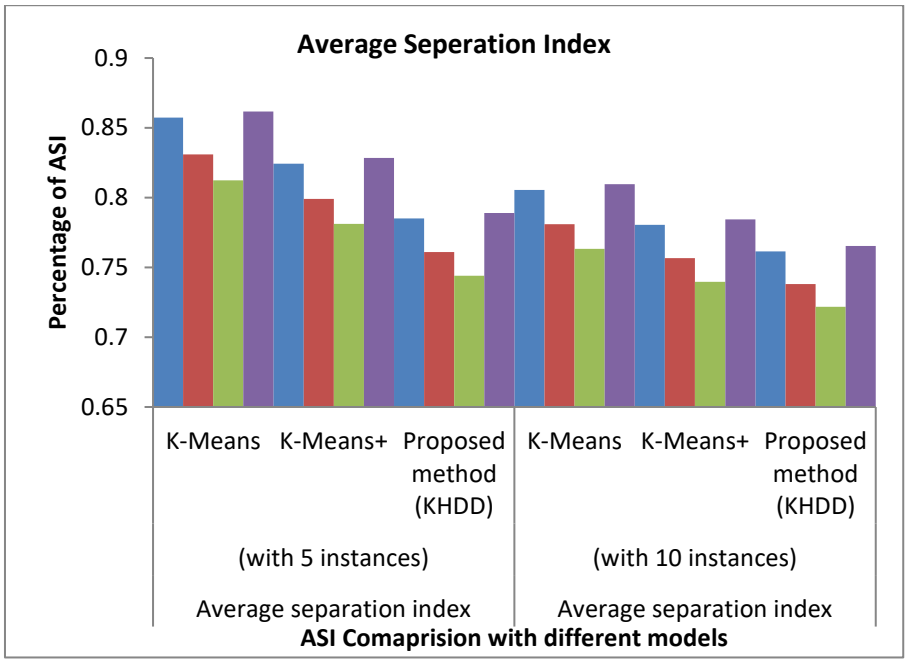

Figure 1. Average Separation Index of the models

Entropy utilizes external data class labels for this situation. Entropy is the difference between the original label of the class and the predicted class label. The lower entropy indicates the better cluster is. The Entropy increases when ground truth of objects in the cluster additionally expands. The more noteworthy entropy implies that the grouping isn't great. So we expect that each group ought to have low entropy to keep up the nature of our grouping. The lower entropy implies better grouping and more noteworthy entropy implies that the grouping isn't great. The proposed algorithm shows that it is having less entropy than other traditional algorithms. Table 2 shows the comparison on different datasets based on entropy between different algorithms and the proposed KHDD model, which had given good results.

Table 2. Comparing different algorithms on different datasets based on Entropy

Data set	Average Entropy %		
	K-Means	K-Means+	Proposed method (KHDD)
Abalone	0.25	0.2375	0.228
Brest cancer	0.29	0.2755	0.26448
Forest fire	0.32	0.304	0.29184
Iris	0.41	0.3895	0.37392

The k-means algorithm is broadly utilized for clustering vast arrangements of information. However, the standard algorithm doesn’t generally ensure great outcomes as the accuracy, which means nearness to original evaluation of the final groups, rely upon the choice of initial centroids.

Figure 2 shows the comparison on different datasets based on entropy between different algorithms and the proposed KHDD model, which had given good results.

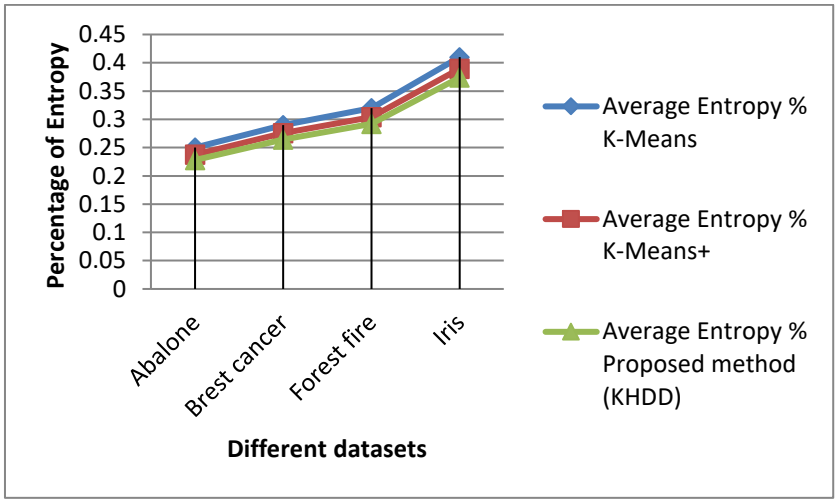

Figure 2. Entropy on different datasets.

Additionally, the computational complexity nature of the typical calculation is frightfully high inferable from the need to reassign the data points various circumstances, amid each iteration of the loop. Table 3 shows the comparison on different datasets based on accuracy between different algorithms and the proposed KHDD model, which had given good results.

Table 3. Comparing different algorithms on different datasets based on Accuracy

Data set	Accuracy%		
	K-Means	K-Means+	Proposed method (KHDD)
Abalone	65	68.3865	72.63330165
Brest cancer	63	66.2823	70.39843083
Forest fire	68	71.5428	75.98560788
Iris	71	74.6991	79.33791411

Sensitivity measures the extent of positives that are effectively recognized. All experiments, regardless of how carefully arranged and executed, have some level of mistake or vulnerability or uncertainty. The accuracy can be evaluated by computing the percentage of the error, which can be calculated when the true value is known. Though the percent error is an absolute value, it can be expressed with magnitude to indicate the direction of error from true value. Table 4 shows the comparison based on true positive, Error and outlier between different algorithms and the proposed KHDD model, which had given good results.

Table 4. Comparing different algorithms on based on Sensitivity and Error rate

Algorithm	True Positive (Sensitivity)	Error %	Outlier %
K-means	72.921	27.91	16.2
K-means+	81.345	22.64	12.4
Proposed method (KHDD)	84.624	19.85	9.92

The Map function splits the contribution to ranges based on the InputFormat and formulates a map job for each assortment in the information. The JobTracker distributes these works to the slave centers. The outcome of each map task is allocated into assembling of key-value sets for each reduce task. The Reduce function at that point collects the different results and consolidates them to answer the bigger problems in data mining that the master core needs to solve. Each reduce task pulls the important parcel from the slave machines where the maps executed, at that point composes its yield once again into HDFS [24]. The proposed algorithm has taken less mapper and reducer time than other traditional algorithm. Table 5 shows the comparison based on mapper time and reducer time on different algorithms and the proposed KHDD model, which had given good results.

Table 5. Comparing different algorithms on based on Map/Reduce time rates

Number of instances	Mapper time(ms)			Reducer time(ms)		
	K-means	K-means+	Proposed method (KHDD)	K-means	K-means+	Proposed method (KHDD)
5	5321	5001.74	4551.5834	5487	5146.806	4842.629765
10	6574	6179.56	5623.3996	6745	6326.81	5952.895529
20	7412	6967.28	6340.2248	7247	6797.686	6395.942757
30	8147	7658.18	6968.9438	8417	7895.146	7428.542871
40	9214	8661.16	7881.6556	9321	8743.098	8226.380908
50	10214	9601.16	8737.0556	10341	9699.858	9126.596392

6. Conclusion and Future Work

Traditional K-means algorithm is the essential for characterizing the information, which is an optimal segment relocation based grouping. It creates information dataset into k-subsets; the value of k is likewise specified by the client. There are some issues with traditional k-means method; problem with initialization of data points, issue of anomalies or outliers, empty groups' issues, high dimensional datasets causes high time complexity nature. In this work, a method has been acquainted which can decrease the computational complexity. The algorithm has been executed for real time collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer for clustering. The proposed method and improved k-means+ has been contrasted and the traditional k-means algorithm. It has been seen

from the yield that the projected algorithm is more efficient than the other two techniques. It plays out the grouping in less time, and is substantially more productive than other methods. There are numerous conceivable outcomes to expand the proposed method in a few fields. This algorithm can be made to work with different real time collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer. The possibility of the participation trust set can be connected to different variations of k-means algorithm for clustering. The most effective method is to select the magnitude or intensity that must be exceeded for a certain reaction, phenomenon, result, or condition to occur or be manifested for a specific dataset is as yet an open question. The speedup of the proposed method might be enhanced much more fundamentally on the off chance that we lessen correspondence time by gathering at least two subsets together on one machine. Be that as it may, this will probably influence scalability of the method. Likewise, our proposed algorithm needs to utilize precisely K machines to work. Future work could make the calculation more adaptable by enabling it to adjust into any number of machines.

References

1. Bikku, Thulasi, N. Sambasiva Rao, Ananda Rao Akepogu. Hadoop based feature selection and decision making models on Big Data. 2016, Indian Journal of Science and Technology, vol. 9, no. 10.
2. Bhatt, Chintan M., S. K. Peddoju, eds. Cloud Computing Systems and Applications in Healthcare. 2016, IGI Global.
3. Queiroz, Rodrigo, et al. Does feature scattering follow power-law distributions? An investigation of five pre-processor-based systems. 2014, Proceedings of the 6th International Workshop on Feature-Oriented Software Development. ACM.
4. Ye, Qiang, Ziqiong Zhang, Rob Law. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. 2009, Expert systems with applications, vol. 36, no. 3, pp. 6527-6535.
5. Endert, A., et al. 2017, The state of the art in integrating machine learning into visual analytics. Computer Graphics Forum.
6. Hartigan, John A., Manchek A. Wong. Algorithm AS 136: A k-means clustering algorithm. 1979, Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no. 1, pp. 100-108.
7. Jain, Anil K. Data clustering: 50 years beyond K-means. 2010, Pattern recognition letters, vol. 31.8, pp. 651-666.

8. Fahim, A. M., et al. An efficient enhanced k-means clustering algorithm. 2006, Journal of Zhejiang University-Science A, vol. 7, no. 10, pp. 1626-1633.
9. Celebi, M. Emre, Hassan A. Kingravi, Patricio A. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. 2013, Expert systems with applications, vol. 40, no. 1, pp. 200-210.
10. Bikku, Thulasi. A Novel Multi-Class Ensemble Model for Classifying Imbalanced Biomedical Datasets. 2017, IOP Conference Series: Materials Science and Engineering, vol. 225, no. 1. IOP Publishing.
11. Arthur, David, Sergei Vassilvitskii. How slow is the k-means method?. 2006, Proceedings of the twenty-second annual symposium on Computational geometry. ACM.
12. Faber, Vance. Clustering and the continuous k-means algorithm. 1994, Los Alamos Science 22.138144.21.
13. Pakhira, Malay K. Clustering large databases in distributed environment. 2009, Advance Computing Conference, 2009. IACC 2009. IEEE International. IEEE.
14. Chen, Dar-Ren, et al. Classification of breast ultrasound images using fractal feature. 2005, Clinical imaging, vol. 29, no. 4, pp. 235-245.
15. Khan, Shehroz S., Amir Ahmad. Cluster center initialization algorithm for K-modes clustering. 2013, Expert Systems with Applications, vol. 40, no. 18, pp. 7444-7456.
16. Nazeer, KA Abdul, SD Madhu Kumar, M. P. Sebastian. Enhancing the k-means clustering algorithm by using a O (n logn) heuristic method for finding better initial centroids. Emerging Applications of Information Technology (EAIT), 2011, 2011 Second International Conference on. IEEE.
17. Wu, Jieming, Wenhu Yu. Optimization and improvement based on K-Means Cluster algorithm. Knowledge Acquisition and Modeling, 2009. KAM'09. Second International Symposium on, vol. 3. IEEE.
18. Mingoti, Sueli A., Joab O. Lima. Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. 2006, European Journal of Operational Research, vol. 174, no. 3, pp. 1742-1759.
19. Topchy, Alexander, Anil K. Jain, William Punch. Combining multiple weak clusterings. Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE.
20. Krzanowski, Wojtek J. Statistical principles and techniques in scientific and social research. 2007, Oxford University Press on Demand.

21. Bikku, Thulasi. A Novel Multi-Class Ensemble Model for Classifying Imbalanced Biomedical Datasets. 2017, IOP Conference Series: Materials Science and Engineering, vol. 225, no. 1. IOP Publishing.
22. Chang, Xiangyu, et al. Sparse Regularization in Fuzzy $ c $-Means for High-Dimensional Data Clustering. 2017, IEEE transactions on cybernetics.
23. Strehl, Alexander, Joydeep Ghosh, Raymond Mooney. Impact of similarity measures on web-page clustering. 2000, Workshop on artificial intelligence for web search (AAAI 2000). Vol. 58.
24. Jung, Se-Hoon, et al. A Novel on Automatic K Value for Efficiency Improvement of K-means Clustering. 2017, Advanced Multimedia and Ubiquitous Engineering. Springer, Singapore, pp. 181-186.