Performance Examination of Hard Clustering Algorithm with Distance Metrics

Simhachalam Boddana, Hymavathi Talla

Abstract*:*Clustering algorithms based on partitions are widely us ed in unsupervised data analysis. K-means algorithm is one the efficient partition based algorithms ascribable to its intelligibility in computational cost. Distance metric has a noteworthy role in the efficiency of any clustering algorithm. In this work, K-means algorithms with three distance metrics, Hausdorff, Chebyshev and cosine distance metrics are implemented from the UC Irvine ml-database on three well-known physical-world data-files, thyroid, wine and liver diagnosis. The classification performance is evaluated and compared based on the clustering output validation and using popular Adjusted Rand and Fowlkes-Mallows indices compared to the repository results. The experimental results reported that the algorithm with Hausdorff distance metric outperforms the algorithm with Chebyshev and cosine distance metrics.

Keywords : Chebyshev distance, cosine distance, distance metrics, Hausdorff distance, k-means.

I. INTRODUCTION

One of the pivotal tasks in data mining is data clustering. Extensive clustering procedures have been proposed for various fields such as pattern identification, image classification, medical diagnostics, Artifitial intelligence etc [7]. The objective of any clustering algorithm is partitioning the information into the desired number of clusters as effectively as possible. Partition based clustering algorithms are quite tolerant in computational performance. Among these algorithms, hard clustering or k-means (KM) clustering algorithm is popular due to its hassle-free implementation and low memory consumption. Clustering methods are implemented by means of an objective function in which some distance metric is used to identify the most similar objects in the data-file.

Distance metric plays a noteworthy role in tweaking any clustering algorithm. Several investigations using distance metrics have been carried out to improve the efficiency of clustering methods.Wen-Liang Hung and Minn-Shen Yang [19] presented a new method based on Hausdorff distance to find the similarity degree between intuitionistic fuzzy sets.

Rucklidge W.J [13] efficiently located objects using the Hausdorff distance. Huttenlocher et. al [6] used Hausdorff metric for comparing images and presented that it can detect the small position errors effectively. Klove Torleiv et. al [9] applied Chebyshev distance metric for constructions of permutation arrays with efficient error correction algorithms. Kahkashan kouser and Sunita [8] compared the performance of the k-means algorithm with three distance metrics, Euclidean, Manhattan and Chebyshev metrics. Lesk AM [10] studied similar substructures extraction in protein sequences by Chebyshev metric. Sung-Hyuk Cha [17] used chebyshev and cosine distance metrics to compare relationships of probability density functions. Strehl et. al [16] studied the cosine distance metric impact on web-page clustering and reported as best metric.Anna Huang [1] investigated the partitional clustering methods' performance for text document datasets with several distance metrics including cosine metric. Najin Dehak et. al [12] used cosine metric for identification of speaker recognition. Sandeep et. al [14] evaluatd the estimation of the selectivity of vector transformation of a string using cosine metric. Steinbach Michael et. al [15] used cosine metric in k-means algorithm to study intra-cluster similarity technique for document clustering techniques.

The classification performance of the k-means algorithm is compared in this paper by taking into account three different distance metrics, namely Hausdorff metric, Chebyshev metric and cosine metric, to three famous physical-world data-files of thyroid, wine, and liver disorder obtained from UCI machine learning. The work primarily focuses on the output result of the clustering method.

The analysis discussed in this paper is as follows: Section 2 outlines the k-means algorithm along with different metrics and the description of the data-files. Section 3 describes experimental outcomes of the process, including discussions, finally the conclusions are recorded in section 4.

II. MATERIALS AND METHODS

Clustering aims at solving to group most similar objects as a cluster in a data-file. Among various clustering algorithms based on partitions, the classical technique algorithm gained much more attention due to its simple computaitonal steps

A. The data-files

The data-files Thyroid, Liver Disease and Wine are the physical-world data-files obtained from the UC Irvine ml-database. The data-file Thyroid is a 5-dimensional data with 215 records. The lab measurements T3(A ratio), amount of T4, total serum T3, amount of TSH and the total change in TSH after hormone injection with respect to the base value are the dimensions and each record called as a specimens.

Revised Manuscript Received on December 13, 2019.

***** Correspondence Author

Simhachalam Boddana*,[1], Department of Mathematics, GITAM University, Visakhapatnam, INDIA. Email: drbschalam@gmail.com

Hymavathi Talla, [1]Department of appied Mathematics, Dr. MRAR PG center, Krishna University, Andhrapradesh INDIA. Email: talla.hymavathianur@gmail.com

The dataset contains 3 distinct groups corresponding to the thyroid activities called Normal containing 150 specimens, Hyperthyroid containing 35 specimens and Hypothyroid containing 30 specimens.

The Liver data-file is a 6-dimensional data with 341 records. Such measurements reflect blood-counts capable of identifying the liver diagnosis that may occur due to excessive consumption of alcohol. These blood-counts are mcv -mean corpuscular volume, alkaline phosphatase, sgpt-alamine aminotransferase, sgot-aspartate aminotransferase, gamma-glutamyl transpeptidase and the amount of alcoholic liquids consumed daily. The 341 specimens were classified into two distinct classes, resulting in 142 specimens in Class 1 and 199 specimens in Class 2 according to the liver diagnosis.

The Wine data-file is a 13-dimensional dataset with 178 specimens.The 13 dimensions reflect the chemical analysis o f the wine made from three different cultivars but grown in t he same Italian area.According to the cultivars, the specimen ss are classified into 3 distinct clusters: Cultivar1 consisting of 59 specimens, Cultivar2 with 71 specimens, and Cultivar3 with 48 specimens.Each sample contains alcohol, malic acid , ash, ash alkalinity, magnesium, total phenols, flavanoids, n onflavanoid phenols, proanthocyanins, color intensity, hue, d iluted wines OD280/OD315 and proline.Author (s) can send paper in the given email address of the journal. There are two email address. It is compulsory to send paper in both email address.

B. Distance Metrics

Any clustering algorithm's effectiveness depends on the di stance metric of the element that calculates the distance betw een objects. A main task of distance metrics is to obtain a suitable similarity function. A metric or distance function is defined as a distance between a set of objects.

Hausdorff distance: Hausdorff distance is also known as Pompeiu-Hausdorff distance [2]. Let $X=\{x_1,x_2\cdots,x_n\}$ and $Y=\{y_1,y_2\cdots,y_k\}$ be two finite point sets, Hausdorff distance metric is denoted by $H(X,Y)$ and defined as $H(X,Y)=\max\{h(X,Y),h(Y,X)\}$ where $h(X,Y)=\max_{x\in X}\min_{y\in Y}\|x-y\|$ and $\|\cdot\|$ is the Euclidean norm on X , Y points. The function $h(X,Y)$ is constituted as directed Hausdorff distance between X and Y . Hausdorff distance calculates the distance from a point in X that is the greatest distance from any point in Y and vice versa and hence results in the degree of mismatch between X and Y .

Chebyshev distance: In a vector space Chebyshev distance or chessboard distance is defined as the maximum distance between two vectors in any coordinate dimension [3]. Chebyshev distance is defined as $d_{ik}=\max_j\{|x_{ij}-y_{kj}|\}$ or in other words it can also be computed the length between two n-dim vectors x in X and y in Y as $dis(x,y)=\max_{i=1,2,\ldots,n}|x_i-y_i|$.

Cosine distance: In an inner product space the cosine distance metric computes the cosine angle between two non-zero vectors. The distance between $x\in X$ and $y\in Y$ is computed as $dis(x,y)=1-\frac{xy'}{\sqrt{(xx')(yy')}}$ .Good quality plagiarism software/ tool (Turnitin / iThenticate) will be used to check similarity that would not be more than 20% including reference section. In the case of exclusion of references, it should be less than 5%.

C. Clustering Indices

Clustering indices are the measure of the similarity between two clustering algorithms. There are several validity indices in the literature. Among those Adjusted Rand Index (ARI) and Fowlkes-Mallows (FMI) indices are more reliable for performance validations. Consider the data-file Z with cardinality $|Z|=N$. Suppose a clustering $C=\{C_1,C_2,\ldots,C_c\}$ such that $|C_i|>0$ partitions the data-file Z . Let $S(Z)$ be the set of all clustering. Let $C'=\{C'_1,C'_2,\ldots,C'_d\}\in S(Z)$ be another one of Z . For C,C' the confusion matrix or contingency table is defined as $M=\lfloor m_{ij}\rfloor_{c\times d}$ where $m_{ij}=|C_i\cap C'_j|$, $1\le i\le c$, $1\le j\le d$.

Adjusted Rand Index: This Adjusted Rand Index (ARI) was proposed by Hubert and Arabie in 1985 [5] and is defined as follows.

$$R_{adj}(C,C')=\frac{\sum_{i=1}^{c}\sum_{j=1}^{d}\binom{m_{ij}}{2}-t_3}{\frac{1}{2}(t_1+t_2)-t_3}$$

Where $t_1=\sum_{i=1}^{c}\binom{|C_i|}{2}$, $t_2=\sum_{j=1}^{d}\binom{|C_j|}{2}$, $t_3=\frac{2t_1t_2}{N(N-1)}$.

For independent clustering the expected index value be zero and one for identical clustering. It also gives negative index values when the numerator in the ratio is negative [18].

Fowlkes-Mallows Index:

Fowlkes and Mallows (FM) [4] developed a validity measure for both hierarchical and flat clustering defined as follows:

$$FM(C,C') = \frac{\sum_{i=1}^{c}\sum_{j=1}^{d} m_{ij}^2 - N}{\sqrt{\left(\sum_i |C_i|^2 - N\right)\left(\sum_j |C'_j|^2 - N\right)}}$$

The index ranged between 0 and 1 where zero indicates independent clustering and one indicates identical clustering.

D. k-means Clustering

The k-means clustering algorithm is one of the efficient partition based algorithms. The method was introduced in 1967 by MacQueen [11]. It is also popular as Hard C-Means algorithm. The method is an iterative method and uses Euclidean distance metric to cluster the data. Consider a dataset Z with N observations or objects to be classified into c (lies between 1 and N) clusters. The method classifies the data in such a fashion that at any one time each observation (object) can only belong to one cluster. Objects can represent as an n -dimensional row vector $z_k = [z_{k1}, z_{k2}, \ldots z_{kn}] \in \Re^n$ and the dataset Z as a matrix of order $N \times n$. The vector of centers is represented by $V = [v_1, v_2, \cdots, v_c]$ where $v_i \in \Re^n$. The rows of the dataset constitute specimens and the columns constitute measurements for these objects. This algorithm has two phases: random selection of the desired number of centers and assigning an object that is close to the center. Since it is an iterative process, the iteration continues until a maximum number of iterations specified or no change in the cluster centers observed. This algorithm converges to local minimum by optimizing the objective function given by

$$J(V) = \sum_{i=1}^{c}\sum_{k=1}^{N} d_{ik},$$

metric d_{ik} is to measure the distance between k^{th} object, z_k and i^{th} centroid, v_i .

The algorithm consists of the steps following:
St 1: Input the number of clusters, c to classify.
St 2: Initialize the centers of the clusters.
St 3: Using a distance metric associate each object with a cluster that has the closest distance from the center of the cluster to the object.
St 4: Update the mean of each cluster as a new centroid
St 5: Apply step 3 and update the clusters
St 7: From St 3 repeat until convergence criterion has been met.

Due to the random process of initializing the cluster cente rs, the algorithm runs a regular number of times to decrease t he sensitivity caused.

III. RESULTS AND DISCUSSION

The algorithm has been executed in ver. R2010a of MATLAB. Authors tested the algorithm with 15 independent test runs until the maximum of 100 iterations or no change in the clusters to achieve best clustering output results.

A. Results

The thyroid data-file that containing 215 specimens is partitioned into three classes. Numbers 1 to 215 label the specimens. The Normal class contains the specimens from 1 to 150, Hyperthyroid class contains the specimens from 150-185 and the specimens from 186 to 215 are in Hypothyroid class. The KM algorithm with Hausdorff, Chebyshev and Cosine distance metrics were implemented to cluster the dataset into three clusters. Using Chebyshev distance metric the algorithm generated three clusters corresponding to clusters Normal with 132 specimens, Hyperthyroid with 45 specimens and Hypothyroid with 38 specimens. 7 Hyper specimens and 4 Hypo specimens were incorrectly allocated to the Normal cluster. 19 Normal specimens and one Hypo sample were wrongly classified into Hyper clusters. 10 Normal specimens and 3 Hyper cluster specimens were incorrectly allocated to the Hypo cluster. The algorithm with Cosine distance metric generates the clusters Normal with 183 specimens, Hyper with 19 specimens and Hypo with 13 specimens. 16 Hyper-related specimenss and 17 Hypo-related specimen s were mistakenly assigned to the Normal Cluster.

The method with Hausdorff distance metric generates three clusters Normal, Hyper and Hypo with 153, 39 and 23 specimens respectively. 10 Hyper-related specimens and 6 Hypo-related specimens are inappropriately assigned to Normal. The Hyper class was inappropriately associated with 13 Normal specimens and one Hypo class sample.

According to the cultivars, the data-file of wine with 178 specimens grouped into three separate groups. The number of the specimenss is 1 to 178. The specimenss fr om 1 to 59 were grouped as cultivar 1, 60 to 130 as cultivar 2 and 131 to 178 as cultivar 3. To classify the data-file with respect to 3 different groups, specifically Cultivar1, Cultivar2 and Cultivar3, the KM algorithm with Hausdorff, Chebyshev and Cosine distance metrics was applied. Three groups corresponding to cultivar1, cultivar2 and cultivar3 with 47, 69 and 62 specimens were produced by KM with Chebyshev metric. The group Cultivar 1 contains one sample of Cultivar 2 and the group Cultivar 2 contains 19 specimens of Cultivar3, 13 Cultivar1 and 20 Cultivar2 specimens are incorrectly grouped into Cultivar3. With 56, 41, and 81 specimens respectively, the three groups Cultivar1, Cultivar2 and Cultivar3 are produced using Cosine metric. Cultivar1 includes 4 Cultivar2 and 2 Cultivar3 specimens in the band, and Cultivar2 includes 6 Cultivar3 specimens. 9 Cultivar1 and 32 Cultivar2 specimenss are grouped incorre ctly in Cultivar3.

By the implementation of Hausdorff metric in KM, it generated three classes as Cultivar1 with 47 specimens, Cultivar2 with 69 specimens and Cultivar3 with 62 specimens. One Cultivar2 sample incorrectly classified as Cultivar1 and 19 Cultivar3 specimens improperly classified as Cultivar2 specimens. 13 Cultivar1 and 20 Cultivar2 specimens inaccurately assigned to Cultivar3.

341 Liver data collection specimens listed as two groups.

The number of the specimens is 1 to 341. Class 1 includes 1 to 142 specimens and Class 2 comprises 143 to 341 specimens. The KM method was applied with Hausdorff, Chebyshev and Cosine distance metrics to cluster the dataset into two clusters, namely Class 1 and Class 2. The dataset was classified in two categories by KM with Che byshev metric, so that Class 1 had 38 specimens and Class 2 had 303 specimens. 25 Class 2 specimens were falsely assigned to Class 1 specimens and 129 Class 1 specimens were falsely assigned to Class 2 specimens. The KM method used Cosine metric produced 57 specimens related with Class 1 & 284 specimens related with Class 2. 41 Class 1 specimens were erroneously assigned and 126 Class 2 specimens were erroneously assigned.

The method KM with Hausdorff metric classified the data-file into two different classes such a way that Class 1 contains 38 specimens and Class 2 contains 303 specimens. 24 Class 2 specimens were assigned to Class 1 and 128 Class 1 specimens were assigned to Class 2 improperly.

Table-I shows the summary results of the clustering method with different distance metrics with the number of specimens satisfactorily and erroneously classified in the corresponding clusters of the data-files

B. Discussion

According to the results obtained by adopting the Hausdorff distance metric in the k-means algorithm for the thyroid dataset, 137 specimens were properly grouped out of 150 normal cluster specimens and the remaining 13 specimens were unfairly assigned to the hyperthyroid cluster. In the 35 specimens of the hyperthyroid cluster, 25 were correctly clustered and the 10 were incorrectly clustered as normal cluster. Given the 30-sample hypothyroid cluster, 23 were properly allocated and 6 specimens were assigned to the normal cluster and one sample was inappropriately assigned to the hyperthyroid cluster. When the algorithm with Chebyshev distance metric is applied out of 150 normal cluster specimens, 121 were correctly classified and 19 specimens were incorrectly classified as hyperthyroid and 10 specimens were incorrectly classified as hypothyroid clusters. Further, out of 35 hyperthyroid specimens, 25 specimens were grouped correctly. 7 specimens were assigned to normal and 3 were assigned to hypothyroid clusters wrongly. Of the 30 hypothyroid cluster specimens, 25 have been correctly identified. One sample has been classified as hyperthyroid and four specimens have been incorrectly identified as a normal group. KM with cosine distance metric 150 specimens of the normal cluster was correctly classified. Out of 35 hyperthyroid specimens, 19 were correctly grouped but 16 specimens were wrongly assigned to cluster normal. Further, out of 30 specimens of hypothyroid 13 were properly grouped and 17 were improperly grouped as normal cluster.

For the thyroid data-file, the k-means algorithm with Hausdorff distance metric achieved an accuracy of about 91.33% for the normal cluster, 71.43% for the hyperthyroid cluster and 83.33% for the hypothyroid cluster. In comparison, the algorithm with Chebyshev metric achieved an accuracy of about 80.67%, 71.43%, 83.33% and with the cosine distance metric achieved an accuracy of about 100%, 54.28%, 43.33% correspondingly.

Based on the outcomes of the k-means technique with Hausdorff distance metric for the wine dataset, 46 specimens were assigned accurately out of 59 specimens of cultivar1 cluster. 13 Cultivar1 specimens were erroneously mapped as cultivar 3 specimens. Whilst using Chebyshev metric, these occurrences are equal to 46 and 13 specimens and 50 and 9 while using cosine metric respectively. Of the 71 cultivar2 specimens, 50 specimens are properly graded. Just one sample mistakenly listed as a sample in cultivar1 & 20 was wrongly identified as cultivar3. When applying Chebyshev metric, these occurrences are equal to 50, 1 & 20 specimens and 35, 4 & 32 when applying cosine metric in KM algorithms respectively. In addition, out of 48 cultivar3 specimens, 29 specimens were properly classified and 19 specimens were incorrectly classified as cultivar2. When Chebyshev metric is applied in KM algorithm, these frequencies are equal to 29 and 19. The KM system with cosine range metric properly allocated 40 specimens to cultivar3 out of 48 and falsely allocated 6 specimens to cultivar2 and 2 specimens to cultivar1.

Table-I: The results acquired by KM cluster algorithm with different distance metrics for Thyroid, Wine and Liver disorder data-files.

KM Clustering Method with		Thyroid data-file (3clusters)			Wine data-file (3clusters)			Liver data-file (2clusters)	
		Normal	Hyperthyroid	Hypothyroid	Cultivar1	Cultivar2	Cultivar3	Class 1	Class 2
Hausdorff metric	Correct	137	25	23	46	50	29	14	175
	Incorrect	16	14	0	1	19	33	24	128

	Total	153	39	23	47	69	62	38	303
Chebyshev metric	Correct	121	25	25	46	50	29	13	174
	Incorrect	11	20	13	1	19	33	25	129
	Total	132	45	38	47	69	62	38	303
Cosine metric	Correct	150	19	13	50	35	40	16	158
	Incorrect	33	0	0	6	6	41	41	126
	Total	183	19	13	56	41	81	57	284

The algorithm k-means with Hausdorff distance metric achieved an accuracy of about 77.96% for cultivar1, 70.42% for cultivar2 and 60.41% for cultivar3 for the wine data-file. In differentiation, KM with Chebyshev metric achieved an accuracy of about 77.96%, 70.42% and 60.41% and with cosine metric achieved an accuracy of about 84.74%, 49.29% and 83.33% correspondingly.

Based on the results of the k-means algorithm with Hausdorff distance metric achieved for the liver disorder data-file, 14 specimens of the clas1 cluster are correctly grouped from 142 specimens. 128 class1 specimens are wrongly listed as Class2 specimens. If Chebyshev metric is applied such occurrences are 13 and 129 specimens respectively if cosine metric is used, such occurrences are equal to 16 and 126 specimens. In addition, 175 specimens are correctly classified out of 199 cluster class2 specimens. 24 class2 specimens were incorrectly classified as class1 specimens. Whilst Chebyshev metric is applied these occurrences are equal to 174 and 25 specimens respectively and while cosine metric is applied, such occurrences are equal to 158 and 41 specimens.

The algorithm k-means with Hausdorff distance metric obtained an correctness of approximately 9.85% and 87.94% corresponding to class1 and class2 for the liver data-file. Compared to Chebyshev metric, the method achieved an accuracy of approximately 9.15 percent and 87.44 percent and with cosine metric, the method reached an accuracy of approximately 11.27 percent and 79.39 percent.

Based on the results obtained for the three distance metrics used in the KM algorithm, Hausdorff distance metric classification output yields its highest with 86.05 percent compared to Chebyshev and cosine metrics yielding 79.53 percent and 84.18 percent respectively for thyroid data-file. Hausdorff, Chebyshev's classification performance and cosine distance metric give its best in the case of wine data-file at 70.22 percent. Hausdorff distance metric's classification performance yields its best at 55.42 percent compared to the Chebyshev and cosine metrics, which yield 54.83 percent and 51.02 percent respectively in the case of data-file for liver disorder.

Table-II summarizes the correctness and performance of the classification in the percentage form of the three metrics.

According to the Adjusted Rand Index obtained for the three distance metrics used in the KM algorithm, Hausdorff distance metric classification performance yields its best of 0.5791 compared to the Chebyshev and cosine metrics yielding 0.4529 and 0.4856 respectively for the thyroid data-file. Hausdorff and Chebyshev distance metrics classification quality is highest at 0.3711 compared to cosine metric, which yields 0.3534 for wine data-file. Hausdorff distance metric classification performance yields the best with -0.0063 compared to Chebyshev metrics and cosine yielding -0.0088 and -0.0137 respectively in the case of data-file for liver disorder.

According to the Fowlkes-Mallows Index obtained for the three distance metrics used in the KM algorithm, Hausdorff distance metric classification performance yields its best performance with 0.8063 compared to the Chebyshev and cosine metrics yielding 0.7216 and 0.8032 respectively for the thyroid data-file. Hausdorff and Chebyshev distance metrics classification quality is highest at 0.5835 compared to cosine metric, which yields 0.5776 for wine data-file. Hausdorff distance metric classification performance yields its best performance with 0.6385 compared to the Chebyshev and cosine metrics that yield 0.6375 and 0.6022 respectively in the case of data-file for liver disorder. Table-III tabulates the clustering validation indexes, the Adjusted Rand Index and the Fowlkes-Mallows Index.

The objective function values corresponding to each iteration are depicted as line graphs in figure 1, figure 2 and figure 3 for the data-files thyroid, wine and liver disorder respectively. The clustering quality of the algorithm KM with Hausdorff, Chebyshev and cosine distance metrics is depicted as a bar graph in figure 4 for the data-files thyroid, wine and liver disorder. Within Figure 4, the x-axis represents the data-files and y-axis represents the percentages of the output of the algorithm.

Table-II: Performance evaluation of the KM clustering algorithm obtained from tests with various distance metrics for data-files of the thyroid, wine and liver disorder.

KM Clustering-Method with	Thyroid data-file (3clusters)				Wine data-file (3clusters)				Liver data-file (2clusters)		
	Correctness %			Classification performance %	Correctness %			Classification performance %	Correctness %		Classification performance %
	Normal	Hyper thyroid	Hypo thyroid		Cultivar 1	Cultivar 2	Cultivar 3		Class 1	Class 2	
Hausdorff metric	91.33	71.43	76.67	86.05	77.96	70.42	60.42	70.22	9.86	87.94	55.42

Chebyshev metric	80.67	71.43	83.33	79.53	77.96	70.42	60.42	70.22	9.15	87.44	54.83
Cosine metric	100	54.28	43.33	84.18	84.74	49.29	83.33	70.22	11.27	79.39	51.02

Table-III: Clustering Performance validation indices of the clustering algorithm KM with different distance metrics for Thyroid, Wine and Liver disorder data-files.

KM Clustering Methods	Thyroid data-file (3clusters) indices		Wine data-file (3clusters) indices		Liver data-file (2clusters) indices	
	Adjusted Rand	Fowlkes-Mallows	Adjusted Rand	Fowlkes-Mallows	Adjusted Rand	Fowlkes-Mallows
Hausdorff metric	0.5791	0.8063	0.3711	0.5835	-0.0063	0.6385
Chebyshev metric	0.4529	0.7216	0.3711	0.5835	-0.0088	0.6375
Cosine metric	0.4856	0.8032	0.3534	0.5776	-0.0137	0.6022

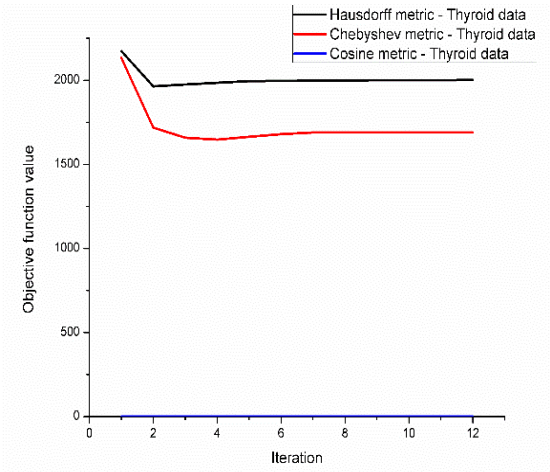

Fig. 1. Objective function values for thyroid data-file

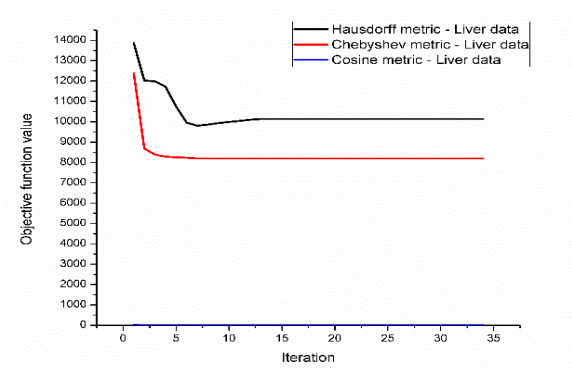

Fig. 3. Objective function values for thyroid data-file

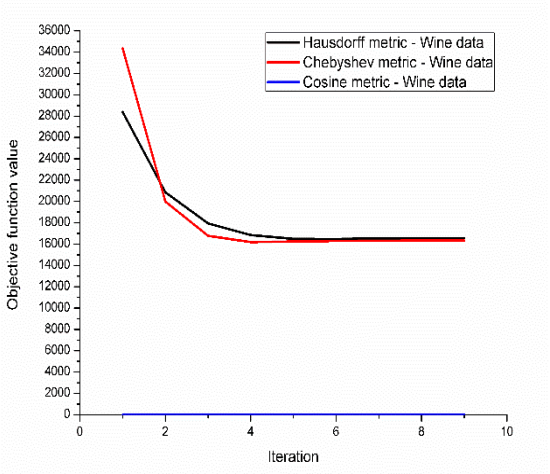

Fig. 2. Objective function values for wine data-file

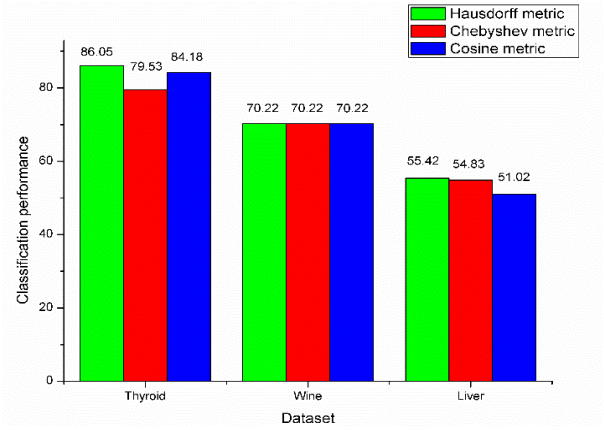

Fig. 4. KM Performance comparison between distance metrics

IV. CONCLUSION

The analysis of results of this work, allow the researchers to analyze the experimental study of k-means algorithm. The k-means algorithm with three different metrics allows tailoring the performance of the algorithm. In this study, the correctness percentage of the clusters using Hausdorff and Chebyshev metrics reports same but differ with cosine and observed that the overall performance of the algorithm with three metrics are similar in the case of wine data-file is

considered. According to the results obtained from the experimental tests using Matlab software, the algorithm with Hausdorff distance metric showed more intense in clustering performance than the distance metrics Chebyshev and cosine. Further, the well-known clustering validation indices, ARI and FM report that the KM algorithm with Hausdorff distance metric performed well. The results are summarized and shown in tabular and graphical formats. As a future work this study can be extended on more diverse data-files.

CONFLICT OF INTEREST

The writers announce that when this paper is published, there is no conflict of interest.

REFERENCES

1. Anna Huang. Similarity Measures for Text Document Clustering. *New Zealand Computer Science Research Student Conference*, April, 2008, pp.49-56.
2. Blumberg and Henry. Hausdorff's Grundzüge der Mengenlehre. *Bulletin of the American Mathematical Society*, 1920, 27 (3): pp. 116–129.
3. Bock R K and Krisher W. *The data analysis brief book*. Springer-Verlag, New York, 1998.
4. Fowlkes E B and Mallows C L. A Method for Comparing two Hierarchical Clusterings. *Journal of the American Statistical Association*, 1983, 78(383): pp.553-569.
5. Hubert L and Arabie P. Comparing Partitions. *Journal of Classification*, 1985, 2: pp.193-218.
6. Huttenlocher, D.P., Klanderman, G.A., and Rucklidge, W.J. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1993, 15(9): pp.850–863.
7. Jain A, Murty M and Flynn P. Data Clustering: A review. *ACM Computing Surveys*, 1999, 31(3): pp. 264-323.
8. Kahkashan kouser and Sunita. A comparative study of K-Means Algorithm by Different Distance Measures. *International Journal of Innovative Research in Computer and Communication Engineering*, 2013, 1(9): pp.2443-2447.
9. Kløve Torleiv, Lin Te-Tsung, Tsai Shi-Cun and Tzeng Wen-Guey. Permutation Arrays Under The Chebyshev Distance. *IEEE Transactions on Information Theory*, 2010, 56(10): pp.2611-2617.
10. Lesk Am. Extraction of geometrically similar substructures: least-squares and Chebyshev fitting and the difference distance matrix. *Proteins*, 1998, 33(3): pp.320-328.
11. MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 1967, vol 1: Statistics: 281-297.
12. Najim Dehak, Reda Dehak, James Glass, Douglas Reynolds and Patrick Kenny. Cosine Similarity Scoring without Score Normalization Techniques. *In Odyssey*: June 2010, pp.15-20.
13. Rucklidge W.J. Efficiently Locating objects using the Hausdorff distance. *International Journal of Computer Vision*, 1997, 24(3): 251-270.
14. Sandeep Tata and Jignesh M Patel. Estimating the Selectivity of tf-idf based Cosine Similarity Predicates. *SIGMOD record*, 2007, 36(2):7-12.
15. Steinbach Michael, George Karypis and Vipin Kumar. A Comparision of Document Clustering Techniques. In *KDD workshop on text mining*, 2000, 400(1): 525-526.
16. Strehl A, Ghosh J, and Mooney R. Impact of similarity measures on web-page clustering. In AAAI-2000: *Workshop on Artificial Intelligence for Web Search*, July 2000, pp.58-64.
17. Sung-Hyuk Cha. Comprehensive Survey on Distance or Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models And Methods In Applied Sciences*, 2007, 1(4):300-307.
18. Wagner Silke and Dorothea Wagner. Comparing Clustering – *An Overview*. *Technical Report*, Faculty of Informatics, University at Karlsruhe(TH), 2007, pp.1-19
19. Wen-Liang Hung and Minn-Shen Yang. Similarity Measures of Intuitionistic Fuzzy sets based on Hausdorff distance. *Pattern Recognization Letters*, 2004, 25(14):1603-1611.

AUTHORS PROFILE

Simhachalam Boddana.
Dr. B. Simhachalam, obtained his M.Sc. (Applied Mathematics, 2005), M.Phil. (Applied Mathematics, 2007) and M.Tech. (Information Technology, 2009) from Andhra University, Visakhapatnam, INDIA. He did his Ph.D. from A.K.N.U. He is presently working as an Assistant Professor in the Department of Mathematics at GITAM University, Visakhapatnam. He is specialized in applied group theory in Mathematics and his area of research interest is soft-computing and data mining.

Hymavathi Talla.
Dr. T. Hymavathi is a distinguished professor in mathematics. She has been teaching for graduate and post graduate students for more than 20 years. She has also guided more than 10 Ph.D. students. She got K.U. II rank in B.Sc., secured I-rank in M.Sc. from R.E.C. Warangal and did her Ph.D. from O.U. She received Samajagan Prathibha Puraskar award and Prof.N.Ch.Pattabhiramacharulu Endowment award. Her area of research interest is in fluid dynamics in Mathematics.