International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 6, June 2017)

K-MEANS Clustering with a Covariance Matrix

Simhachalam B[1], Hymavathi T[2]

[1]Department of Mathematics, GIT, GITAM University, Visakhapatnam, Andhra Pradesh 530045, India
[2]Department of Mathematics, Adikavi Nannaya University, Rajahmundry, Andhra Pradesh 533296, India

Abstract **- The performance of clustering algorithms desperately depends on a metric defined over the input space. K-means technique is a vigorous partition based clustering algorithm in the datamining discipline. In this article covariance matrix based distance metric is proposed to improve the classification performance of the k-means algorithm. The metric alleviates overfitting, distortion in the shapes of clusters issues that occur in standard k-means algorithm. The empirical study on three well-known real-world datasets from UC Irvine repository and the results of popular Adjusted Rand and Fowlkes-Mallows indices demonstrate that the proposed metric causes better efficient clustering and hence it can be employed for notably ameliorate in the clustering performance.**

Keywords **- k-means, Euclidean distance, covariance matrix, distance metric, hard clustering.**

I. INTRODUCTION

Clustrering defines categorization of indistinguishable objects. K-means (KM) or hard clustering technique is a robust partion based clustering technique with a numerous applications in diverse fields including pattern recognition, medical sciences, web clustering, animals and plants, psychiarty and earth sciences [11][13]. Its main feature is optimizing the squared error criterion [2][18].

K-means clustering technique employs Euclidean distance metric which primarly aims to reduce the within-cluster distances. Despite the wide use of this metric, there are some drawbacks that have led to improvement of efficency of the algorithm. Since the metric is used to compute the mean and variance which might cause the sensitivity in estimating the outliers. It produces only spherical type clusters inspite of the hidden structures [4][16]. Since the algorithm is very sensitive to the initialization due to its local optimization many researchers have been proposed several initialization methods to imporve the efficency. Chen Zhang and Shixiong Xia [5] introduced the concept of sub-merger to combined the categories instead of random initialiazation of centers. Fahim et. al [8] implemented the technique to imporve the efficency of the algorithm by storing the data of preceding iteration to be used in the succeeding iteration. In this implementation the computed distances from the data point to clusters' centers are stored.

In the next iteration instead of computing the distance between the data point to all prototypes, it calculates the distance from the data point to the nearest cluster prototype which was known in the preivious iteration for the datapoint may be prone to changing the cluster. Nazeer and Sebastian [1] introduced two phases: determination of initial prototypes and assigninig the ojects to the cluster. In the first phase the data set is paritioned in to k (number of clusters) sets where each set contains the data points that are having nearest distance among them and then computes the initial centroids. In later phase the approach of creating clusters is similar to the Fahim's approach.

Rather than emplyoing the enhancements in initialization techniques for k-means method many researchers propsed metric based imporvement techniques. Bin Zhang [3] propsed harmaonic means as a distance metric in k-means algorithm that to handle outliers. Bin quoted that the metric is a dynamic weighting function because a data point that is not nearby any cluster was given a high weight so that it could participate in the next clustering iteration. In each iteration the metric function automatically regulate. Matousek [15] implemented the k-means algorithm by introducing a cost function with non-linear approximation for any given fixed positive epsilon. Tapas Kanungo et. al [20] introduced local approximation technique in k-means algorithm on swaping prototypes in order to improve practical performance. Tapas Kanungo et. al [21] propsed an algorithm called filtering algorithm. A kd-tree structure was introduced in the k-means method to improve the algorithms' runtime.

This article presents a new distance metric by evaluating a cluster covariance matrix and distance-inducing matrix. This metric has the capable of handling different geometical shapes that inherent in the data. The classification performance of the k-means algorithm is analyzed with the proposed distance metric and compared to standard k-means algorithm by considering three different famous real-world data sets thyroid, wine and liver disorder obtained from UC Irvine machine learning repository. The work primarily focuses on the cluster output result of the clustering method

The work presented in this paper is as follows: The k-means algorithm, proposed distance metric and the details of the datasets are given in brief in section 2. In section 3 emperical study of the method including discussions are presented and the conclusions are reported in section 4.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 6, June 2017)

II. Materials And Methods

A. The Dataset

In this experimental study the famous real-world data sets Thyroid, Liver Disorder and Wine datasets are used. The respective data sets are donated by Danny Coomans [6], Richard [17] and Forina [9] and obtained from the UC Irvine Machine Learning Repository. The Thyroid data set contains 215 samples where each sample has 5 lab measurements, Liver data set consists of 341 samples where each sample has 6 types of blood tests that have the potential of recognizing the liver disorders which might arise due to excessive alcohol consumption and the Wine data set contains 178 samples where each sample is characterized by 13 types of chemical analysis of the wine derived from three different cultivars but grown in the same region in Italy. The Thyroid dataset contains three different groups according to the thyroid functions namely Normal with 150 samples, Hyperthyroid with 35 samples and Hypothyroid with 30 samples. According to the liver disorders, the 341 samples are grouped into two distinct classes such that 142 samples in Class 1 and 199 samples in Class 2. The samples are clustered into three different clusters as stated by the cultivars: Cultivar 1 consisting of 59 samples, Cultivar 2 with 71 samples and Cultivar 3 containing 48 samples. The attributes of these data sets are shown in table I.

TABLE I
THE ATTRIBUTES OF THE THYROID, LIVER DISORDER AND WINE DATA SETS.

Data set	Attributes
Thyroid	1. T3-resin uptake test (A percentage). 2. Value of total serum thyroxine given by the isotropic displacement method. 3. Total serum triiodothyronine value given by radioimmunoassay. 4. Value of basal thyroid-stimulating hormone (TSH) given by radioimmunoassay. 5. After injection of 200 micrograms of thyrotropin-releasing hormone the maximal absolute difference of TSH value as compared to the basal value.
Liver disorder	1. Mean corpuscular volume (mcv). 2. Alkaline phosphotase (alkphos). 3. Alamine aminotransferase (sgpt). 4. Aspartate aminotransferase (sgot). 5. Gamma-glutamyl transpeptidase (gammagt). 6. The number of half-pint equivalents of alcoholic beverages drunk per day (drinks).
Wine	1. Alcohol 2. Malic acid. 3. Ash. 4. Alcalinity of ash. 5. Magnesium. 6. Total phenols. 7. Flavanoids. 8. Nonflavanoid henols. 9. Proanthocyanins. 10. Color intensity. 11. Hue. 12. OD280/OD315 of diluted wines. 13. Proline.

B. K-means clustering algorithm

Clustering aims at partitioning the data objects into coherent groups. In clustering algorithms, k-means algorithm is more prominent since its ease of execution, computational reliability and less memory utilization [19] [22]. In 1967 MacQueen [14] proposed an iterative based clustering algorithm known as k-means method. It is also popular as Hard c-means algorithm. This algorithm be composed of two main activities. The first activity is the initialization of determined number of cluster prototypes (centers) and the second is assignment of the data points to the closest prototype. In this method the interspace between data point and the prototype is determined by Euclidean distance. Let us consider a dataset Z with N observations or objects to categorize into c $(1 \le c \le N)$ clusters. The method partitions the data in a manner that at any one time each observation (object) can only belong to one cluster and in each iteration the prototypes are updated by calculating the mean of the data points in the cluster. the data points 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 $N \times n$ matrix. The vector of prototypes is represented by $V = [v_1, v_2, \cdots, v_c]$ where $v_i \in \Re^n$. The rows of the dataset represent objects or samples or data points and the columns represent characteristics of the data points. The algorithm iterates continuously until it meets a specific criterion like unchanging in the cluster prototypes or maximum number of iterations are performed. The convergence of the algorithm is achieved by minimizing the objective functional given by

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

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

The algorithm comprises the following fundamental steps.

Step 1: Chose the number of clusters, c.
Step 2: Assign initial prototypes.
Step 3: Assign data points to clusters that having closest interspace between the cluster's prototype and the point by using a distance metric.
Step 4: Evaluate the mean of each cluster as update the prototype and clusters
Step 5: Iterate step 3 and step 4 until convergence criterion has been met.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 6, June 2017)

The sensitivity of the method occurs due to the random initialization of cluster prototypes. For obtaining better results and to diminish the sensitivity the algorithm run a frequent number of times.

C. Distance metrics

Distance metrics influence the performance of any clustering algorithm. Defining a good metric is an important aspect in producing better clustering results.

Euclidean distance: Consider a n-dimensional vector space. The Euclidean distance between two vectors $X=(x_1,x_2,x_3,\ldots,x_n)$ and $Y=(y_1,y_2,y_3,\ldots,y_n)$ is defined as follows:

$$d(X,Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}\ .$$

The Euclidean distance between the vectors z_k and z_i of the data matrix Z is evaluated as follows:

$$d_{ik}=\sqrt{(z_k-z_i)^T(z_k-z_i)},\quad 1\le i,k\le N\ .$$

For Z the resulted Euclidean distance matrix is of $N\times N$ ordered symmetric matrix that contains zeros as its principal diagonal. The Euclidean distance metric fabricates only spherical shaped clusters.

Proposed distance: Employing the sample mean in standard k-means algorithm distorts the shapes of the clusters. This causes the sensitivity of outliers. The feature of local optimization in k-means algorithm affects the structure of the cluster inspired to propose the distance metric by evaluating the cluster covariance matrix. The eigen values and eigen vectors of the covariance matrix elucidate the inherent geometrical structures and orientation of the clusters [7]. The distance metric is defined as follows:

$$d_{ikA_i}=\sqrt{(z_k-v_i)^T A_i(z_k-v_i)},\ 1\le k\le N,\ 1\le i\le c \quad (4)$$

.

In equation (4) replacing $A_i=I$ gives Euclidean distance; here A_i is considered by an norm-inducing matrix. The clusters with different geometrical structures will ensue using this distance norm in one input space. For each cluster the metric is evaluated by within cluster norm-inducing matrix, which is used for optimization of the objective function (1).

The expression for A_i is defined as

$$A_i=\left[\rho_i\det(F_i)\right]^{n^{-1}}F_i^{-1},\ 1\le i\le c \quad (5)$$

Where the i^{th} cluster's fuzzy covariance matrix F_i is given by $F_i=\sum_{k=1}^{N}(z_k-v_i)(z_k-v_i)^T,1\le i\le c$.

The weighting parameter $\rho_i>0$ is introduced for obtaining a feasible solution and constrained as $|A_i|=\rho_i$ (volume constraint). The eigen values and eigen vectors describe the geometrical structures of the clusters and since we are dealing with matrices, when an eigen value is zero then the matrix may become singular so that the calculation of inverse in equation (5) could not be possible. This can be overcome by constraining the maximal and minimal eigenvalues ratio. It is achieved by introducing a threshold β given as follows: $\lambda_{ij}=\max_j\lambda_{ij}/\beta$ for all j for which $\max_j\lambda_{ij}/\lambda_{ij}>\beta$. When the threshold constrain fails i.e. when the ratio exceeds the threshold the minimum eigenvalue is raised to satisfy the constrain and then covariance matrix is revamped by: $F=\Phi\mathrm{E}\Phi^{-1}$ where E is a diagonal matrix contains eigen values and the matrix Φ constitutes the corresponding eigenvectors as columns. In equation (5) F_i is revamped as $F_i=[\phi_{i1}\ldots\phi_{in}]\,diag(\lambda_{i1}\ldots\lambda_{in})[\phi_{i1}\ldots\phi_{in}]^{-1}$.

(2)

(3)

Further, the problem of overfitting may occur when a cluster contains very smaller amount of data points. This leads to maximum elongation of the cluster. To overcome this drawback, the covariance matrix is revamped by adding a scaled identity matrix as given below.

$$F_i=(1-\gamma)F_i+\gamma\det(F_0)^{1/n}I$$

Where $0\le\gamma\le1$ is a tuning parameter and F_0 is the data set's covariance matrix. When the clusters' number grows, then it effects the shape of the clusters being in circular shape so that the value of the tuning parameter is used to describe the shape of the cluster.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 6, June 2017)

D. Guideline for parameter tuning

There are three parameters ρ, β and γ in the proposed distance metric. Since the objective function (1) is linear in A_i, it cannot be optimized with respect to A_i. To obtain feasible solution the parameter ρ is introduced in equation (5). The value of ρ determines the volume of the cluster. Without a priori one can fix ρ at 1 for each cluster. The threshold β is introduced for avoiding the matrix become singularity. If the maximal and minimal eigen values ratio is extremely high (say 10^{20}) the matrix is almost singular and hence inverse calculation in equation (5) is not possible. The tuning parameter γ in equation (6) is incorporated for the shapes of the clusters. When $\gamma=1$ all the covariance matrices are same with equal size. To avoid the dependency on the number of clusters F is constructed using whole data set. The term $\det(F_0)^{1/n}$ is incorporated to diminish the tuning effort involved. In application ρ (for volume), β (for singularity) and γ (for shapes) can be optimized in cross-validation runs and the optimal values corresponding to highest accuracy will be chosen.

E. Clustering indices

Clustering indices are the measure of the similarity between two clustering algorithms. In the literature among various validity indices Adjusted Rand Index (ARI) and Fowlkes-Mallows (FMI) indices are more favorable for performance validations. Consider the data set Z with cardinality $|Z|=N$. Suppose the data set Z is partitioned by a clustering $C=\{C_1,C_2,\ldots,C_c\}$ such that $|C_i|>0$. Let $C'=\{C'_1,C'_2,\ldots,C'_d\}\in S(Z)$ be another clustering of Z where $S(Z)$ be a clustering's set. The confusion matrix or contingency table of the pair C,C' 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: Hubert and Arabie proposed Adjusted Rand Index (ARI) in 1985 [12]. This index is defined by

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

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 [23].

Fowlkes-Mallows Index:

A validity index called Fowlkes-Mallows Index (FM) was introduced by Fowlkes and Mallows [10] for both hierarchical and flat clustering and is 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.

III. Results And Discussion

The software MATLAB version R2010a is used to implement the algorithm with both the distance metrics. 15 independent test runs were tested with the criterion that until the maximum of 100 iterations or no change in the clusters is occur to accomplish better clustering. For the experimental study the parameters are considered as follows: $\rho_i=1$, $\forall i$, $\beta=10^{15}$, $\gamma=1$ for thyroid data set and $\gamma=0.3$ for wine and liver disorder data sets.

A. Results

The thyroid data set consists of 215 samples is classified in to three classes. These samples are numbered from 1 to 215. The first 150 samples belong to Normal class, from 151 to 185 samples belong to Hyperthyroid class and the remaining belong to Hypothyroid class.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 6, June 2017)

The standard KM algorithm creates three clusters, normal with 151 samples, hyperthyroid with 33 samples and hypothyroid with 31 samples. 7 samples of hyperthyroid class and 7 samples of hypothyroid class have been wrongly assigned to normal cluster. 8 samples of normal cluster are improperly grouped as hyperthyroid cluster. The hypothyroid cluster included 5 normal cluster samples and 3 hyperthyroid samples. The KM algorithm with the proposed metric generates three clusters, normal with 153 samples, hyperthyroid with 39 samples and hypothyroid with 23 samples. 10 samples that belong to hyperthyroid class and 6 samples that associated with hypothyroid class are misassigned to normal class. 13 samples of normal class and one sample of hypothyroid class are mistakenly grouped as hyperthyroid class. None of the normal and hyperthyroid samples are grouped in hypothyroid class.

According to the cultivars the wine data set containing 178 samples is categorized in to three different cultivars. The samples are labelled by numbers 1 to 178. cultivar 1 consists of first 59 samples, from 60 to 130 samples belong to cultivar 2 and the cultivar 3 has remaining samples. The standard KM algorithm clustered the data set into three different clusters named as cultivar 1, cultivar 2 and cultivar 3 consisting 47, 69 and 62 samples respectively. In cultivar 1 only one sample of cultivar 2 is assigned wrongly and 19 samples of cultivar 3 is incorrectly assigned to cultivar 2. 13 samples of cultivar 1 and 20 samples of cultivar 2 are wrongly grouped in to cultivar 3 cluster. With the implementation of proposed distance metric, the KM method clustered the data set in to three different clusters, cultivar 1 with 80 samples, cultivar 2 with 50 samples and cultivar 3 with 48 samples. The cluster cultivar 1 contains 20 samples of cultivar 2 and one sample of cultivar 3. 50 samples of cultivar 2 are properly clustered in to cultivar 2 and none was assigned wrongly. One sample that belong to cultivar 2 is improperly assigned to cultivar 3.

The 341 liver disorder data set samples are numbered from 1 to 341 and. The data set has two classes. The first 142 samples belong to class 1 and class 2 contains remaining samples. The standard KM technique classified the data set in to two classes such that class 1 contains 38 samples and class 2 contains 303 samples. Class 1 contains 24 class 2 samples and class 2 contains 128 class 1 samples inappropriately. The KM technique with the proposed metric produces two classes, class 1 and class 2 having 95 and 246 samples respectively. 43 samples of class 2 inappropriately assigned to class 1 and 90 class 1 samples are assigned to class 2 incorrectly.

Table II shows the results summary of the clustering method with different distance metrics with number of samples that are classified appropriately and inappropriately into the respected clusters of the data sets.

B. Discussions

From the results obtained by the implementation of standard k-means algorithm for the thyroid data set among 150 samples of normal class 137 are properly grouped. In the remaining 13 samples, 8 samples are assigned to hyperthyroid class and 5 samples are assigned to hypothyroid class improperly. Out of 35 hyperthyroid class samples 25 are correctly assigned. From the remaining 10 samples, 7 samples grouped to normal class and 3 are grouped to hypothyroid class incorrectly. Only 7 samples of hypothyroid class are wrongly assigned to normal class. The method with the proposed distance metric results out of 150 normal class samples 137 are correctly classified. 13 samples are incorrectly classified as hyperthyroid class. Similarly, out to 35 hyperthyroid class samples 25 are truly assigned and faulty assignment of 10 samples to normal class. In 30 samples of hypothyroid 23 samples are properly grouped. 6 samples are grouped to normal class and one sample is grouped to hyperthyroid class incorrectly.

For the thyroid dataset, k-means clustering with proposed metric attained an accuracy of about 91.33% for normal class, 71.43% for hyperthyroid class and 76.67% for hypothyroid class. In comparison, with Euclidean metric, the method attained an accuracy of about 91.33%, 71.43% and 76.67% respectively.

The results obtained from the standard KM method for wine data set demonstrate that out of 59 cultivar 1 samples, 46 are correctly classified and the remaining are incorrectly classified as cultivar 3 samples. For 71 cultivar 2 samples, 50 samples are appropriately associated to cultivar 2. One sample is assigned to cultivar 1 and 20 samples are assigned to cultivar 3 inappropriately. Out of 48 samples of cultivar 3, 29 samples are grouped correctly and the rest are incorrectly grouped as cultivar 2 samples. With the proposed metric all the 59 samples of cultivar 1 are flawlessly classified. Out of 71 samples of cultivar 2, 50 samples are rightly assigned to cultivar 2. One sample is assigned to cultivar 3 and 20 are assigned to cultivar 1 incorrectly. 47 samples out of 48 samples are flawlessly grouped as cultivar 3 but only one is improperly assigned to cultivar 1.

For the dataset wine, k-means clustering with proposed metric attained an accuracy of about 100% for cultivar 1, 70.42% for cultivar 2 and 97.92% for cultivar 3.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 6, June 2017)

In comparison, with Euclidean metric, the method attained an accuracy of about 77.96%, 70.42% and 60.41% respectively.

For the liver disorder dataset, according to the obtained results of the k-means technique with Euclidean metric, for 142 samples of class 1, 14 samples are correctly classified. 128 samples are incorrectly classified as class 2 samples. These frequencies are equal to 52 and 90 in case of proposed metric is applied. Further, for 199 samples that belong to class 2, 175 are appropriately assigned to class 2 and 24 samples are inappropriately classified as class 1 samples. These frequencies are equal to 156 and 43 when the proposed metric is implemented.

For the liver dataset, the method k-means with proposed metric attained an accuracy of about 36.62% for class 1 cluster and 78.39% for class 2 cluster. In comparison, with Euclidean metric, the method attained an accuracy of about 9.85% and 87.94% respectively.

The error-free and the classification performance in percentage form of the two metrics are summarized in table III.

According to the Adjusted Rand Index obtained for the proposed and Euclidean distance metrics used in KM algorithm the clustering performance of both metrics reports similar performance with 0.5791 in case of thyroid data set. The clustering performance of proposed distance metric gives its best with 0.6592 compared to Euclidean metric which yield 0.3711 in case of wine data set. In case of liver disorder data set the clustering performance of proposed distance metric yields its best with 0.0409 comparing to the metric Euclidean which yield -0.0063

According to the Fowlkes-Mallows Index obtained for the proposed and Euclidean distance metrics used in KM algorithm the clustering performance of both metrics reports similar performance with 0.8063 in case of thyroid data set. The clustering performance of proposed distance metric gives its best with 0.7765 compared to Euclidean metric which yield 0.5835 in case of wine data set. In case of liver disorder data set the clustering performance of proposed distance metric yields its best with 0.5715 comparing to the metric Euclidean which yield 0.6385 The clustering validation indices, Adjusted Rand Index and Fowlkes-Mallows Index are tabulated in table IV.

TABLE II
THE CLUSTERING RESULTS OBTAINED BY KM ALGORITHM WITH DIFFERENT DISTANCE METRICS FOR THYROID, WINE AND LIVER DISORDER DATA SETS.

KM Clustering Method with		Thyroid data set (3 clusters)			Wine data set (3 clusters)			Liver data set (2clusters)	
		Normal	Hyperthyroid	Hypothyroid	Cultivar 1	Cultivar 2	Cultivar 3	Class 1	Class 2
Proposed metric	Correct	137	25	23	59	50	47	52	156
	Incorrect	16	14	0	21	0	1	43	90
	Total	153	39	23	80	50	48	95	246
Euclidean metric	Correct	137	25	23	46	50	29	14	175
	Incorrect	14	8	8	1	19	33	24	128
	Total	151	33	31	47	69	62	38	303

TABLE III
PERFORMANCE EVALUATION OF THE CLUSTERING ALGORITHM KM OBTAINED BY THE RESULTS WITH DIFFERENT DISTANCE METRICS FOR THYROID, WINE AND LIVER DISORDER DATA SETS.

KM Clustering Method with	Thyroid data set (3 clusters)				Wine data set (3 clusters)				Liver data set (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	
Proposed metric	91.33	71.43	76.67	86.05	100	70.42	97.92	87.64	36.62	78.39	60.99
Euclidean metric	91.33	71.43	76.67	86.05	77.96	70.42	60.41	70.22	9.85	87.94	55.42

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 6, June 2017)

TABLE IV
CLUSTERING PERFORMANCE VALIDATION INDICES OF THE CLUSTERING ALGORITHM KM WITH PROPOSED AND EUCLIDEAN DISTANCE METRICS FOR THYROID, WINE AND LIVER DISORDER DATA SETS.

KM Clustering Methods	Thyroid data set (3 clusters)		Wine data set (3 clusters)		Liver data set (2clusters)	
	Adjusted Rand Index	Fowlkes-Mallows Index	Adjusted Rand Index	Fowlkes-Mallows Index	Adjusted Rand Index	Fowlkes-Mallows Index
Proposed metric	0.5791	0.8063	0.6592	0.7765	0.0409	0.5715
Euclidean metric	0.5791	0.8063	0.3711	0.5835	-0.0063	0.6385

The objective function values corresponding to each iteration are depicted as line graphs in figure 1, figure 2 and figure 3 for the data sets thyroid, wine and liver disorder respectively. The classification performance of the algorithm KM with Euclidean and proposed distance metrics is depicted as a bar graph in figure 4 for the data sets thyroid, wine and liver disorder. In figure 4, the x-axis represents the data sets and the performance percentages of the algorithm are represented by y-axis.

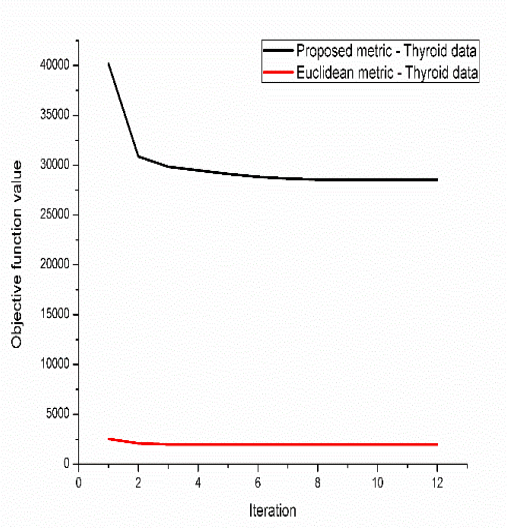

Figure 1. Objective function values for thyroid data set

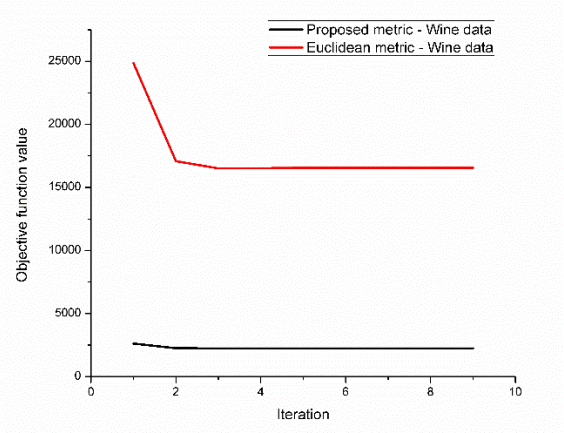

Figure 1. Objective function values for wine data set

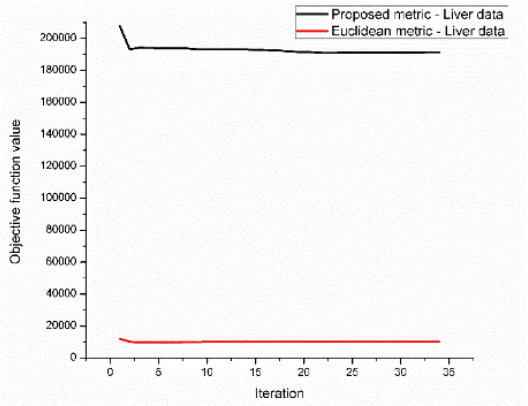

Figure 2. Objective function values for liver disorder data set

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 6, June 2017)

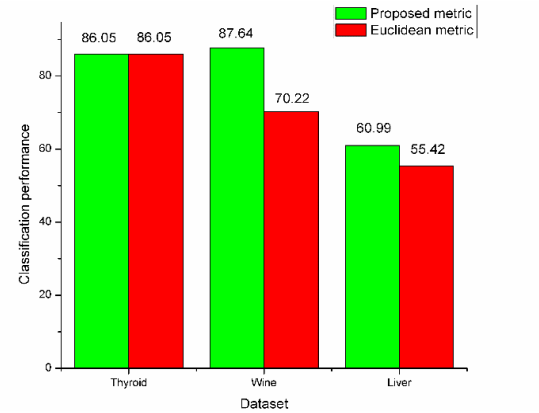

Figure 4. KM Performance comparison between distance metrics

IV. CONCLUSION

Distance metric plays a significant role in tweaking any clustering algorithm. An extensive research investigations using distance metrics have been carried out to improve the efficiency of clustering methods. In this article a distance metric relied on covariance matrix is proposed. This metric employes local norm inducing matrix so that the clusters with dissimilar structres that inherent in a data set can be shaped. The risks that occur of being singularity matrix or by considering linearly correlated or homogeneous data have been dealt with parameters. Although the computational cost is slightly high, the metric can effectively handles outliers in producing clusters. The emprical study on the UCI data sets demonstrated that the proposed metric can be an aid to imporve clustering performance. The popular indices ARI and FM values report that the algorithm KM with proposed distance metric preformed well. The results are summarized and shown in tabular and graphical formats. As a future study, the proposed technique can be extended on more different data sets.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

REFERENCES

[1] Abdul Nazeer K. A and Sebastian M. P., "Improving the accuracy and efficiency of the k-means clustering algorithm", In International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009), 1:308-312, July 2009, London, UK.

[2] Anderberg M. R., "Cluster analysis for applications", Academic Press Inc., London, 1973.

[3] Bin Zhang, "Generalized k-harmonic means - dynamic weighting of data in unsupervised learning", In proceedings of the 2001 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics:1-13; April 2001, Chicago, USA.

[4] Bradley P. S and Fayyad U.M., "Refining initial points for K-Means clustering", In Proc. 15th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA: 91–99, 1998.

[5] Chen Zhang and Shixiong Xia, " K-means Clustering Algorithm with Improved Initial center," In Second International Workshop on Knowledge Discovery and Data Mining (WKDD), Moscow: 790-792, 2009.

[6] Coomans, I. Broeckaert, M. Jonckheer and D. L Massart, "Comparison of Multivariate Discrimination Techniques for Clinical Data – Application to the Thyroid Functional State", Methods of Information in Medicine, 22: 93-101, 1983.

[7] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan and Stuart Russell, "Distance metric learning, with application to clustering with side-information", Advances in Neural Informataion Processing System, 15(1): 505-512, 2002.

[8] Fahim M, Salem A. M, Torkey F. A and Ramadan M. A, "An Efficient enhanced k-means clustering algorithm", Journal of Zhejiang University, 10(7): 1626-1633, 2006.

[9] Forina M, Stefan Aeberhard and Riccardo Leardi, (1991). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, Genoa, Italy

[10] Fowlkes E B and Mallows C L, "A Method for Comparing two Hierarchical Clusterings", Journal of the American Statistical Association, 78(383): 553-569, 1983.

[11] Hartigan, J. A., "Clustering Algorithms", Wiley, New York, 1975.

[12] Hubert L and Arabie P, "Comparing Partitions", Journal of Classification, 2: 193-218, 1985.

[13] Jain A, Murty M and Flynn P, "Data Clustering: A review", ACM Computing Surveys, vol.31, no.3, pp. 264-323. 1999.

[14] MacQueen J, "Some Methods for Classification and Analysis of Multivariate Observations". Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, vol 1: Statistics: 281-297, 1967.

[15] Matousek J, "On Approximate Geometric *k*-Clustering", Discrete and Computational Geometry, 24(1): 61-84, 2000.

[16] Pena J. M, Lozano J. A and Larranaga P, "An empirical comparison of four initialization methods for the k-means algorithm", Pattern Recognition Letters, 20(10): 1027-1040, 1999.

[17] Richard S. Forsyth, (1990). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Mapperley Park, Nottingham NG3 5DX, England.

[18] Salton G, "Automatic Text Processing", Addison-Wesley, New York, 1989.

[19] Steinbach Michael, George Karypis and Vipin Kumar, "A Comparision of Document Clustering Techniques", In KDD workshop on text mining, 400(1): 525-526, 2000.

[20] Tapas kanango, David M Mount, Nathan S Netanyahu, Christine D Paitko, Ruth Silverman and Angela Y Wu, "A Local Search Approximation Algorithm For K-Means Clustering", Computational Geometry, 28(2,3): 89-112, 2004.

[21] Tapas kanango, David M Mount, Nathan S Netanyahu, Christine D Paitko, Ruth Silverman and Angela Y Wu, "An efficient k-means clustering algorithm: analysis and implementation", IEEE Transcations on Pattern Analysis and Machine Intelligence, 24(7):881-892, 2002.

[22] Trevor Hastie, Robert Tibshirani and Jerome Friedman, "The Elements of Statistical Learning", Springer-Verlag, 2008.

[23] Wagner Silke and Dorothea Wagner, "Comparing Clustering – An Overview", Technical Report, Faculty of Informatics, University at Karlsruhe(TH), 1-19, 2007.