International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (03), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Comparative study of Symmetry Based Automatic data clustering techniques: *Application to edge pixels clustering of digital images*

Partha Ghosh[1], Kalyani Mali[2], Sitansu Kumar Das[3]

[1]*Computer Sc. and Engineering, Govt. College of Engineering and Ceramic Technology (India)*

[2]*Computer Sc. and Engineering, University of Kalyani (India)*

[3]*Computer Science, Chittaranjan College (India)*

ABSTRACT

Several different clustering algorithms have been proposed to deal with clusters with various geometric shapes. Those algorithms can detect compact clusters, straight lines, shells, and contours with polygonal boundaries or well-separated non-convex clusters. One thing that should be highlighted is that there is no clustering algorithm which can tackle all kinds of clusters. In this paper, the evolutionary clustering technique is described that uses the new line symmetry based distance measure. At first the traditional K-Means algorithm is described that uses the simple yet effective mean-based distance measure. Then we proceed on with the point symmetry based distance measure called SBKM (Symmetry Based K-Means Algorithm) and then we discuss the line symmetry based distance measure and their results on artificial and real data sets. The mentioned algorithms are all applicable to unsupervised clustering paradigms. Our first objective is to determine automatically the optimal number of clusters in any data set. Second, it attempts to find clusters of arbitrary shapes and sizes. We show that line symmetry based distance can give very promising results, without a priori knowledge of the actual number of clusters, if applied to the automatic clustering problem. We have compared the line symmetry based distance algorithm with two other clustering techniques: K-Means and SBKM.

Keywords: K-means, SBKM, Rand index, DB index, Line symmetry, Clustering.

I.INTRODUCTION

Clustering is one of the most common unsupervised data mining methods to explore the hidden structures embedded in a data set. In Supervised Learning: the data we feed our algorithm is "tagged" to help our logic make decisions. Example: Bayes spam filtering, where we have to flag an item as spam to refine the results but in Unsupervised Learning: it tries to find correlations without any external inputs other than the raw data. Example: data-mining clustering algorithms.

Clustering gives rise to a variety of information granules whose use reveals a structure of data. In order to identify clusters mathematically in a data set, it is usually necessary to first define a measure of similarity or proximity. This measure will allow us to assign data points to a cluster i.e., assign patterns to the domain of a

particular cluster center. One commonly used measure of similarity is the Euclidean distance, D between two patterns $\bar{x}$ and $\bar{z}$ defined by $D = ||\bar{x} - \bar{z}||$. Smaller Euclidean distance means better similarity and vice-versa. Symmetryis considered as a pre-attentive feature that enhances recognition and reconstruction of shapes and objects. The word Pre-attentive: Derived from Pre-attention that is noticing of something before attention is fully focused on it. The exact mathematical definition of symmetry (Symmetry is a type of invariance: The property that something doesn't change under a set of transformations) is inadequate to describe and quantify symmetry found in the natural world or those found in the visual world.

It is reasonable to assume that some kind of symmetry occur in the structures of clusters, because symmetry is so common in the abstract and in the nature. But the immediate problem is how to measure symmetry.Zabrodsky et al [1] have proposed a kind of symmetry distance to detect symmetry in a figure extracted from an image. Their basic strategy is to choose the symmetry that is closest to the figure measured by an appropriated measure. Here the minimum sum of the squared distances (Euclidean distance) over which the vertices must be removed to impose the assumed symmetry is adopted. The goal of clustering is to reduce the amount of data by categorizing or grouping similar data items together.

II.RELATED WORK

A. K-MEANS CLUSTERING ALGORITHM

K-Means is one of the simplest unsupervised learning algorithms, which is used when we have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. So, its aims to find the positions of the clusters that minimize the square of the distance from the data points to the cluster center. Finally, it targets at minimizing an objective function. The objective function is given by Eq. 1.

$$J(V) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}(\|x_i - v_j\|)^2 \qquad (1)$$

Where, $\|x_i - v_j\|$ is the Euclidean distance between x_i andv_j, 'c_i'is the number of data points in i^{th} cluster, 'c' is the number of cluster centers.Let $X = \{x_1, x_2, x_3, \ldots\ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, v_3, \ldots\ldots\ldots, v_c\}$ be the set of centers. Randomly select 'c' cluster centers. Calculate the distance between the data point and cluster centers. Assign the data point to the cluster center whose distance from the cluster center is the minimum of all the centers. Recalculate the new cluster center using Eq. 2.

$$v_i = (1/c_i)\sum_{j=1}^{c_i} x_j \qquad (2)$$

Where, C_i represents the number of data points in the i^{th} cluster. Recalculate the distance between each data point and the new obtained cluster centers.If no data point was reassigned then stop, otherwise repeat these process.

Advantages:Fast, robust and easier to understand. Relatively efficient: $O(tknd)$, where n is no. of objects, k is no. of clusters, d is no. of dimension of each object, and t is no. of iterations. Normally $k, t, d << n$, gives best

result when data set are distinct or well separated from each other.Disadvantages:The learning algorithm requires a priori specification of the number of cluster centers. If there are two highly overlapping data then K-means will not be able to resolve that there are two clusters. The learning algorithm is not invariant to non-linear transformations (i.e., with different representations of data we get different results). Euclidean distance measures can unequally weight underlying factors. It is unable to handle noisy data and outliers. Algorithm fails for non-linear data set and easy to get stuck at the local optimal solutions.

B. SBKM Algorithm

In order to improve the performance of the K-means algorithm, several improved K-means algorithms have been developed in the past several years. A new type of non-metric distance, based on point symmetry, is proposed by Su and Chou [2] which is used in a K-means based clustering algorithm, referred to as symmetry based K-means (SBKM) algorithm. SBKM is found to provide good performance on different types of data sets where the clusters have internal symmetry. However, it can be shown that SBKM will fail for some data sets where the clusters themselves are symmetrical with respect to some intermediate point since the point symmetry distance ignores the Euclidean distance in its computation.

It has been mentioned in a subsequent paper by Chou et al [3] where they have suggested a modification, the modified measure has the same limitation of the previous one [2]. No experimental results have been provided in [3]. In order to overcome the limitation of being easy to get stuck at the local optimal solutions, (which is a drawback of the regular K-means clustering algorithm), some attempts have been made to use genetic algorithms for clustering data sets [4-6]. To overcome the problem of automatic cluster determination from the data sets. Recently, many automatic clustering techniques have been introduced. These automatic clustering techniques are based on genetic algorithm methods and Differential Evolution (DE) methods.

Handl and Knowles [7] proposed multi-objective clustering with automatic *K*-determination (MOCK) to detect the optimal number of clusters from data sets. But due to the heuristic nature of the algorithm, it provides an approximation to the real (unknown) Pareto front only. Saha and Bandyopadhyay [8] proposed a multi objective clustering technique. In this algorithm points are assigned to different clusters based on the point symmetry based distance. It is able to detect clusters having point symmetry property. However it will fail for clusters having nonsymmetrical shape.

Most clustering algorithms assume the number of clusters to be known a priori. The desired granularity [9] is generally determined by external, problem criteria. There seems to be no definite answer to how many clusters are in data set a user defined criterion for the resolution has to be given instead. Second, most of the existing clustering algorithms adopt 2-norm distance measures in the clustering. These measures fail when clusters tend to develop along principal axes. The symmetry based clustering techniques also seek for clusters which are symmetric with respect to their centers. Thus, these techniques will fail if the clusters do not have this property.

- ***Point Symmetry Based Distance:***

Symmetry is considered as a pre-attentive feature that enhances recognition and reconstruction of shapes and objects. Su and Chou [2] presented an efficient point symmetry distance (PSD) measure to help partitioning the

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (03), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

data set into clusters where each cluster has the point symmetry property. Given Npatterns $x_j, (j = 1, \ldots, N)$, and a reference vector c(e.g., a cluster centroid), the point symmetry distance (PSD) between a pattern $x_j$and the reference vector $c$is defined by Eq. 3.

$$d_s(\bar{x}_j, \bar{c}) = \min_{i=1,\ldots,N \text{ and } i \neq j} \frac{\|(\bar{x}_j - \bar{c}) + (\bar{x}_i - \bar{c})\|}{\|(\bar{x}_j - \bar{c})\| + \|(\bar{x}_i - \bar{c})\|} \qquad (3)$$

Where the denominator term is used to normalize the distance so as to make it insensible to the Euclidean distances $||(\bar{x}_j - \bar{c})||$ and $||(\bar{x}_i - \bar{c})||$. It may be noted that the numerator of the equation is actually the distance between the mirror image point of $\bar{x}_j$with respect to $\bar{c}$ and its nearest neighbor in the data set. If the right hand term of the above equation is minimized when $\bar{x}_i = \bar{x}_{j^*}$ then the pattern $\bar{x}_{j^*}$is denoted as the symmetrical pattern relative to $\bar{x}_j$with respect to c. Here it can be easily seen that the above equation is minimized when the pattern$\bar{x}_i = (2\bar{c} - \bar{x}_j)$exists in the data set $(i.e., d_s(\bar{x}_j, \bar{c}) = 0)$.Based on this point symmetry based distance, the algorithm was proposed that mimics the K-Means algorithm but assigns the patterns to a particular cluster depending on the symmetry based distance $d_s$rather than the Euclidean distance, only when d_s is greater than some user specified threshold [2], $\theta = 0.18$. Otherwise assignment is done according to the Euclidean distance, as in normal K-means.

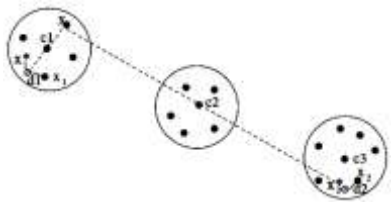

Figure 1

In the above Figure 1, we have three clusters that are well separated. The centers of these clusters are denoted by $\bar{c_1}, \bar{c_2}, \bar{c_3}$, respectively. Let us take the point$\bar{x}$. After the application of K-means algorithm, point $\bar{x}$ is being assigned to the cluster 1. But when SBKM is applied on the result given by K-means algorithm, the following will happen. The symmetrical point of $\bar{x}$ with respect to $\bar{c}_1$ is $\bar{x}_1$. Since it is the first nearest neighbor of the point $\bar{x_1^*} = (2 \times \bar{c_1} - \bar{x})$. Let the Euclidean distance between $\bar{x_1^*}$and $\bar{x_1}$be d_1. So the symmetrical distance of $\bar{x}$ with respect to $\bar{c_1}$is given by Eq. 4.

$$d_s(\bar{x}, \bar{c_1}) = d_1 / (d_e(\bar{x}, \bar{c_1}) + d_e(\bar{x_1}, \bar{c_1})) \qquad (4)$$

Here $d_e(\bar{x}, \bar{c_1})$ and $d_e(\bar{x_1}, \bar{c_1})$ are the Euclidean distances of $\bar{x}$ and $\bar{x_1}$ from $\bar{c}_1$, respectively. Similarly, symmetrical point of $\bar{x}$ with respect to $\bar{c}_2$ is$\bar{x_2}$. And the symmetrical distance of $\bar{x}$ with respect to $\bar{c}_2$ is given by Eq. 5.

$$d_s(\bar{x}, \bar{c_2}) = d_2 / (d_e(\bar{x}, \bar{c_2}) + d_e(\bar{x_2}, \bar{c_2})) \qquad (5)$$

Let $d_2 < d_1$; and obviously the denominator term of Eq. 4 is less than the denominator term of Eq. 5, because the Euclidean distance between $\bar{x}$ and $\bar{c}_2$ and the Euclidean distance between $\bar{x_2}$ (its symmetrical point) and $\bar{c}_2$ is so much larger than the Euclidean distance between $\bar{x}$ and $\bar{c}_1$ and the Euclidean distance between $\bar{x_1}$ (its symmetrical point) and $\bar{c}_1$. Therefore $d_s(\bar{x}, \bar{c_1}) \gg d_s(\bar{x}, \bar{c_2})$ and $\bar{x}$ is assigned to $\bar{c}_2$. This will happen for

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (03), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

other points as well finally resulting in merging of the three clusters after application of SBKM.Su and Chou have chosen θ equals to 0.18. However we have observed that clustering performance is significantly affected by the choice of θ and its best value is dependent on the data characteristics. No guideline for the choice of θ is provided in [2]. To overcome the limitations, a new definition of the point symmetry based distance was proposed [10].

- ***A New Definition of Point Symmetry Distance:***

The PS-based distances, $d_s$will fail when the clusters themselves are symmetrical with respect to some intermediate point. It has been shown, in such cases the points are assigned to the farthest cluster. In order to overcome this limitation, [10] proposed a new PS distance which is called $d_{ps}(\bar{x}, \bar{c})$associated with point $\bar{x}$ with respect to a center $\bar{c}$. The proposed point symmetry distance is defined as follows: let a point be $\bar{x}$. The symmetrical (reflected) point of $\bar{x}$ with respect to a particular center $\bar{c}$ is $2 * \bar{c} - \bar{x}$. Let us denoting this by $\bar{x}^*$. Let the first and the second unique nearest neighbors of $\bar{x}^*$ be at Euclidean distances of d_1 and d_2, respectively. Then $d_{ps}(\bar{x}, \bar{c})$ is given by Eq. 6

$$d_{ps}(\bar{x}, \bar{c}) = \frac{(d_1+d_2)}{2} \times d_e(\bar{x}, \bar{c}) \qquad (6)$$

Where, $d_e(\bar{x}, \bar{c})$ is the Euclidean distance between the point $\bar{x}$ and $\bar{c}$. The basic differences between the PS-based distance in [2] and the point symmetry distance$d_{ps}(\bar{x}, \bar{c})$, in [10] are as follows:

1. Instead of finding the Euclidean distance between the original reflected point $\bar{x}^*$ and its first nearest neighbor in [2], here the average distance between $\bar{x}^*$ and its first and second unique nearest neighbor have been taken. Consequently, the term here: $\frac{(d_1+d_2)}{2}$ will never be equal to 0, and the effect of $d_e(\bar{x}, \bar{c})$, the Euclidean distance, will always be considered. This will reduce the problems talk over in Figure 1.
2. Considering both $d_1$and d_2 in the computation of d_{ps}makes the PS-distance more robust and noise resistant.
3. A rough guideline of the choice of θ, the threshold value on the PS-distance is also provided in [10]. It is to be noted that if a point is indeed symmetric with respect to some cluster center then the symmetrical distance computed in the above way will be small, and can be bounded. Let d_{NN}^{max} be the maximum nearest neighbor distance in the data set can be represented by Eq. 7.

$$d_{NN}^{max} = \max_{i=1,\dots,N} d_{NN}(\bar{x}_i) \qquad (7)$$

Where, $d_{NN}(\bar{x}_i)$ is the nearest neighbor distance of$\bar{x}_i$.Assuming that $\bar{x}^*$ lies within the data space, it may be noted that:

$$d_1 \leq \frac{d_{NN}^{max}}{2} \qquad (8)$$

Thus, the threshold θ equals d_{NN}^{max}. Hence for N points and K clusters, the time complexity of assigning the points to the different clusters is O (N^2K).

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (03), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

2.1PROPOSED WORK

C. Existing Line Symmetry Based Distance

What is line symmetry? For a 2-dimenstional figure, if it can be folded in such a way that one-half of it lies exactly on the other half is said to have line symmetry. The idea of line symmetry is very clear and simple but an immediate problem is how to find a metric to measure line symmetry. A kind of line symmetry distance was proposed in [11-13].In this approach, the symmetrical line of a data set is defined by a center vector and an angle between the major axis of the data set and the x-axis. The information of the major axis of the data points belonging to a class or a cluster is computed by the moment of order $(p + q)$ method. Then the major axis is treated as the symmetrical line of that class or cluster.

Saha and Maulik [14] proposed new line symmetry based automatic genetic clustering technique called variable string length genetic line symmetry distance based clustering (VGALS-Clustering). To measure amount of line symmetry of a point $x$with respect to a particular line$i$, $d_{ls}(x, i)$, the following steps are followed:

1. For a particular data point x, calculate the projected point p_i on the relevant symmetrical line i.
2. Find $d_{sym}(x, p_i)$using Eq. 9.

$$d_{sym}(x, p_i) = \frac{\sum_{i=1}^{k} d_i}{k} \quad (9)$$

Where k-nearest neighbors of $x^* = (2 * p_i - x)$ are at Euclidean distances ofd_i, $i = 1, 2, \ldots, k$. Then the amount of line symmetry of a particular point x with respect to that particular symmetrical line of cluster i is calculated using Eq. 10.

$$d_{ls}(x, i) = d_{sym}(x, p_i) \times d_e(x, c) \quad (10)$$

Where c is the centroid of the particular cluster i and $d_e(x, c)$ is the Euclidean distance between the points x and c.

But a problem may exist in this line symmetry measure. This is called lacking of closure property and this would result in a poor clustering. The closure property can be expressed as follows: if the data point p_i is currently assigned to a cluster centroid c_k in the current iteration, the determined most symmetrical point p_j relative to c_k must have been assigned to c_k in the previous iteration. To overcome this problem we have described the line symmetry measure in different way.

D. The Newly Proposed Line Symmetry Based Distance measure:

Given a particular dataset, we first find the symmetrical line of each cluster by using the central moment technique [15]. Let the data set is denoted by $X = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, then the $(p, q)^{th}$ order moment is defined by Eq. 11.

$$m_{pq} = \sum_{\forall (x_i, y_i) \in X} x_i^p y_i^q \quad (11)$$

This is basically derived from the moment formula represented by Eq. 12.

$$m_{p,q} = \iint x^p y^q f(x, y)\, dx dy \quad (12)$$

Moments are generally classified by the order of the moments. The order of a moment depends on the indices p and q of the moment, $m_{p,q}$ and vice versa. Order of the moment $m_{p,q} = p + q$ (sum of the indices)

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (03), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Considering this, the 1st order moments ((p, q) = (1, 0) or (0, 1)) are given by Eq. 13 and Eq. 14.

$$m_{1,0} = \iint x f(x,y)\, dxdy \quad (13)$$

$$m_{0,1} = \iint y f(x,y)\, dxdy \quad (14)$$

The first order moments contain information about the center of gravity of the object given by Eq. 15 and Eq. 16.

$$x_c = \frac{m_{1,0}}{m_{0,0}} \quad (15)$$

$$y_c = \frac{m_{0,1}}{m_{0,0}} \quad (16)$$

The centroid of the given data set for one cluster is defined as $\left(\frac{m_{1,0}}{m_{0,0}}, \frac{m_{0,1}}{m_{0,0}}\right)$. From the spatial moments the *central moments* can be derived by reducing the spatial moments with thecenter of gravity (x_c, y_c) of the object. The central moment is defined by Eq. 17.

$$u_{pq} = \sum_{\forall (x_i y_i) \in X} (x_i - \bar{x})^p (y_i - \bar{y})^q \quad (17)$$

Here $\bar{x} = \frac{m_{1,0}}{m_{0,0}}$ and $\bar{y} = \frac{m_{0,1}}{m_{0,0}}$. According to the calculated centroid and the Eq. 17, the major axis of each cluster can be determined by the following two items:

1. The major axis of the cluster must pass through the centroid.
2. The angle between the major axis and the x-axis is equal to $0.5 \times \tan^{-1}(2 \times u_{11}/u_{20} - u_{02})$

Here, central moment of 2nd order is used in the computation. Thus, we see that for one cluster, its analogous major axis is represented by Eq. 18.

$$\left(\left(\frac{m_{1,0}}{m_{0,0}}, \frac{m_{0,1}}{m_{0,0}}\right), 0.5 \times \tan^{-1}(2 \times u_{11}/u_{20} - u_{02})\right) \quad (18)$$

The acquired major axis is treated as the symmetric line of the related cluster.In order to measure the amount of line symmetry of a point (x_i) w.r.t. a particular line k of cluster C_k, $d_{ls}(x_i, C_k)$ the following steps are followed.

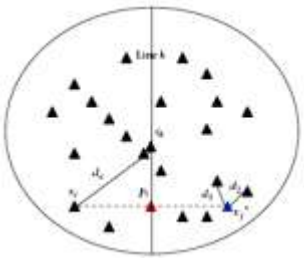

Figure 2

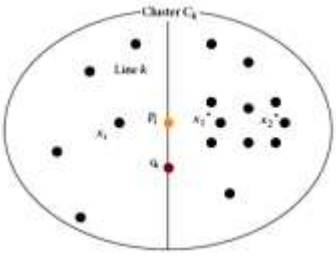

Figure 3

1. (As in Figure 2) for a particular data point x_i, calculate the projected point p_i on the appropriate symmetrical line k of the cluster C_k and then find out all the possible symmetrical data points x_j relative to each symmetrical line k for $1 \le i \le n, 1 \le j \le n,$ and $1 \le k \le K$.
2. Find $d_{sym}(x_i, p_i)$ by Eq. (19).

$$d_{sym}(x_i, p_i) = \frac{\sum_{i=1}^{knear} d_i}{knear} \quad (19)$$

Where k nearest neighbors of $x_j^*$and they are at Euclidean distances of d_i, $i = 1,2,\dots,knear$. The parameter $Knear$ can be set by the user based on specific knowledge of the application. In general, a fixed value of $Knear$ may have many drawbacks. For clusters with too few points, the points likely to be scattered and the distance between two neighbors may be too large. For very large cluster fixed number of neighbors may not be enough because few neighbors would have a distance close to zero. $Knear$ should be much smaller than the number of objects in the data. To gain a clear idea of the distance of the neighborhood of a point, we have chosen $knear \leq \sqrt{n}$ in our implementation. The amount of line symmetry of a particular point $x_i$with respect to particular symmetrical line of cluster C_k is calculated by Eq. 20.

$$d_{ls}(x_i, C_k) = d_{sym}(x_i, p_i) \times d_e(x_i, c_k) \qquad (20)$$

Here, c_k is the centroid of the cluster $C_k$and $d_e(x_i, c_k)$ is the Euclidean distance between data point $x_i$and cluster center c_k.

Now to satisfy the closure property in our proposed line symmetry distance measure, we have to impose some constraint. To compute the line symmetry distance of the data point x_i, we have restricted the candidate symmetrical points $x_j \notin C_k$ relative to each symmetrical line k of the corresponding cluster C_k. For the data point x_i relative to symmetrical line of cluster C_k, this restriction can help us to search more suitable symmetrical point x_j, because we ignore the candidate most symmetrical point x_j which is not in the cluster C_k. We applied the second modification in which the first and second symmetrical points x_1^* and x_2^* of point x_i are found in cluster C_k (as in Figure 3) relative to the symmetrical line, not in all data points; that is, each point x_i, $1 \leq i \leq n$, is assigned to cluster C_k iff $d_{ls}(x_i, C_k) \leq d_{ls}(x_i, C_j)$,where $j, k = 1,\dots,k$ and $j \neq k$, $d_{ls}(x_i, C_k)/d_e(x_i, c_k) \leq \theta$,and x_1^* and x_2^* belong to cluster C_k. The distance $d_{ls}(x_i, C_k)$ is calculated by Eq. 20, and $\theta = d_{NN}^{max}$ is the symmetrical threshold, where $d_{NN}^{max} = \max_{i=1,\dots,N} d_{NN}(\bar{x_i})$ and the distance $d_{NN}(\bar{x_i})$ is the maximum nearest neighbor distance in the data set. The value of θ is kept equal to the maximum nearest neighbor distance among all the points in the data set. Point assignment based on proposed line symmetry distance is given in **Algorithm A.**

Algorithm A: Clustering based on proposed line symmetry distance.

- **Assignment of data points**:

for(i=1;i<=n;i++)

{

for(k=1;i<=K;i++)

{

Find the first and the second symmetrical points x_1^* and x_2^* of x_i relative to a projected point p_i on line k of cluster C_k /*to ensure the closure property */

Calculate the line symmetry-based distance $d_{ls}(x_i, C_k)$, $k = 1, 2, \dots, K$ by Eq. 20.

} /* end of inner for */

Find $C_k = Arg \min_{k=1,\dots,K} d_{ls}(x_i, C_k)$

if($d_{ls}(x_i, C_k) \leq d_{ls}(x_i, C_j)$) /* where $k, j = 1, \ldots, K$ and $k \neq j$ */

{

Assign the point x_i to the cluster C_k. Provided that $d_{ls}(x_i, C_k)/d_e(x_i, c_k) \leq \theta$

}

Else

{

Assign the point x_i to the cluster C_k based on the Euclidean distance measure, $C_k = Arg\ \min_{k=1,\ldots,K} d_e(x_i, c_k)$.

}

} /* end of outer for */

- **Updation of centres**: Compute new cluster centers of the *K* clusters by: $c_k^{new} = \frac{1}{n_k}\sum_{i \in S_k} x^i$. Where, n_k is the number of data points belonging to the cluster C_k and S_k is set of data points which have been assigned to the cluster center c_k.

2.3 Experimental Results and Comparative Study

E. Evaluation of Clustering Quality

The algorithms are implemented in JAVA and are tested on artificial and real data sets. The results of 2 dimensional data sets are displayed and easy to compare and analyze. The qualities of clustering results are measured by adjusted Rand index [16]. *i.e.,* to compare the performance of algorithms (K-Means, SBKM and Proposed Algorithm) adjusted Rand index technique is used. Adjusted Rand index is limited to the interval [0, 1] with a value of 1 with a perfect clustering. The high value of adjusted Rand index indicates the good quality of clustering result. The average and standard deviation of adjusted Rand index for data sets produced by K-Means, SBKM and Proposed Algorithm are depicted in Tables 1(a) and 1(b), respectively.

From these results we can say that, the point symmetry based algorithm is supposed to be an improvement over the traditional K-Means algorithm and similarly the line symmetry based algorithm is supposed to be an improvement over the point symmetry based algorithm.

F. Results on Artificial Data Sets

Data set-1: This data set consists of two bands as shown in Figure 4(a), where each band consists of 200 data points. The final clustering results achieved after application of K-means, SBKM and Proposed Algorithm are shown in Figures 4(b), 4(c) and 4(d) respectively. Proposed algorithm is able to find out the proper clustering for this data. As expected *K*-means and SBKM shows poor performance for this data since the clusters are not hyper-spherical in nature. Our proposed algorithm is able to detect the proper partitioning from this data set as the clusters possess the line symmetry property.

Data set-2: This data set contains 400 points distributed on two crossed ellipsoidal shells shown in Figure 5(a). The final results corresponding to *K*-means, SBKM and Proposed Algorithm are shown in Figures 5(b), 5(c) and

5(d) respectively.. As expected *K*-means and SBKM are not able to detect the proper partitioning but Proposed Algorithm is able to do so.

Data Set-3: This data set is a combination of ring shaped, compact and linear clusters, as shown in Figure 6(a). The clustering result achieved by the *K*-means algorithm is shown in Figure 6(b). The final clustering result of the SBKM algorithm is shown in Figure 6(c). Figure 6(d) shows that the proposed algorithm works well for a set of clusters of different geometrical structures. Both K-means and SBKM clustering algorithms provide $K = 3$ number of clusters in different runs but both are unable to perform the proper partitioning from this data set. Proposed clustering algorithm detects $K = 3$ the optimal number of clusters and the proper partitioning in all consecutive runs.

G. Results on Real Data Sets

The real data sets are taken from UCI repository (http://archive.ics.uci.edu/ml/index.php). For experimental results two real data sets are considered.

(1) Iris: As seen from Table 1(a), the adjusted Rand index of Proposed Algorithm is the best for Iris, while the performance of SBKM is second. However, it can be seen from Tables 1(a) and 1(b) that the performance of K-Means algorithm is found poor. K-Means, SBKM and Proposed Algorithm provide $K = 3$ as the appropriate number of clusters form this data set in all successive runs.

(2) Wine: From Tables 1(a) and 1(b), it is obvious that Proposed Algorithm performs the best for this data set. The adjusted Rand index value achieved by Proposed Algorithm is also the maximum (mention Table 1(a)).

III.APPLICATION: EDGE PIXELS CLUSTERING OF DIGITAL IMAGE

Most of the natural scenes, such as leaves of plants, have the line symmetry property. Figure 7(b-d) shows the two real leaves. First the sobel edge detector [15] is used to find the edge pixels (edge maps) in the input data points which are shown in Figure 8(a-d). The clustering result achieved after execution of the K-means, SBKM and Proposed Algorithm are shown in Figure 9. The proposed algorithm shows a satisfactory clustering result.So, the color image is first converted to black and white and then the edge maps (edge pixels) are obtained using the sobel edge detection [15] technique. Following that, the clustering algorithms namely, traditional K-Means, SBKM and Proposed Algorithm are used to estimate the clusters.

After running the *K*-means algorithm, the obtained clustering is shown in Figure 9(a, d, g, j). After running the SBKM algorithm, the obtained clustering is shown in Figure 9(b, e, h, k). After running the proposed algorithm, the obtained clustering is shown in Figure 9(c, f, i, m).

H. Evaluation Of Clustering Quality

To compare the performance of all three algorithms (K-Means, SBKM and Line-symmetry based proposed algorithm), Davies-Bouldin (DB) index [17] is used. The Davies-Bouldin index is an internal evaluation scheme and is defined as the ratio of inter scatter to intra-scatter distances. Smaller values for DB index correspond to good clusters. That is better the separation of the clusters and “tightness” inside the clusters.Once again from

Figure 10 we found that the proposed algorithm outperformed the Ordinary k-Means algorithm and the SBKM algorithm.

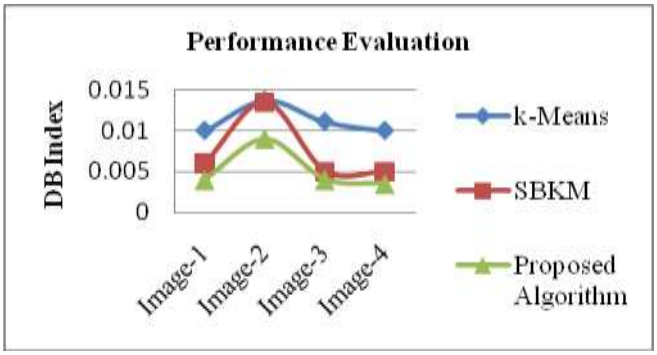

Figure 10

VI. DISCUSSION AND CONCLUSION

Future work includes defining some other form of symmetry such as plane symmetry etc. Developing some clustering techniques based on the proposed line symmetry distance is another direction of future work. The randomized K-d trees based nearest neighbor search can be used to reduce the computation time of the nearest neighbors search mechanism. Instead of using a single straight line in this algorithm, we can try to incorporate curved line/lines to achieve better results. Other than the clustering experiments using leaf example, it is an interesting future research topic to extend the results of this paper to the detection of symmetrical objects in digital images. Current work is going on to improve the proposed clustering technique.

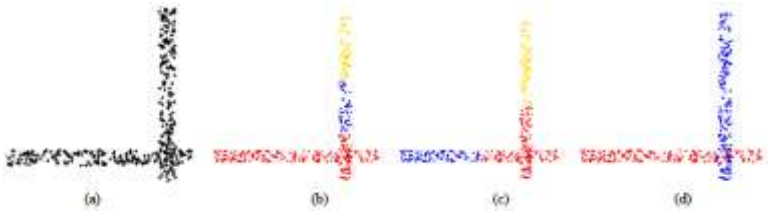

Figure 4: (a)Data set-1 and clustering results achieved after application of (b)K-means, (c)SBKM and (d)Proposed Algorithm.

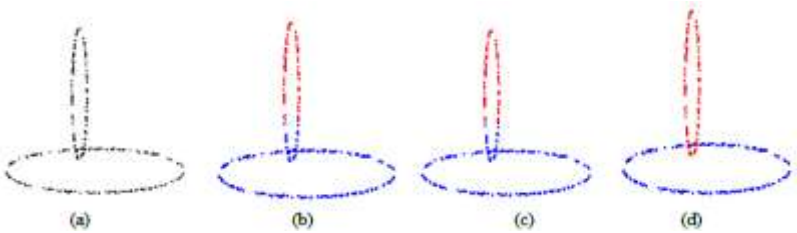

Figure 5: (a)Data set-2 and clustering results achieved after application of (b)K-means, (c)SBKM and (d)Proposed Algorithm.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (03), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

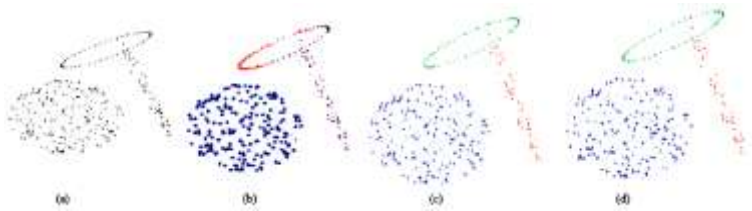

Figure 6: (a)Data set-3 and clustering results achieved after application of (b)K-means, (c)SBKM and (d)Proposed Algorithm.

Table 1: (a) Average of adjusted Rand index for K-means, SBKM and Proposed Algorithm (b) Standard deviation of adjusted Rand index for K-means, SBKM and Proposed Algorithm.

Table 1: (a) Data sets	Number of points (N)	Number of dimensions	Number of clusters (K)	Average value of adjusted Rand index		
				K-Means	SBKM	Proposed Algorithm
Data set-1	400	2	2	0.7467	0.9750	0.9786
Data set-2	400	2	2	0.7585	0.9830	0.9845
Data set-3	350	2	3	0.7491	0.9245	0.9424
Iris	150	4	3	0.7575	0.9240	0.9780
Wine	178	13	3	0.6471	0.9485	0.9567
Table 1: (b) Data sets	Number of points (N)	Number of dimensions	Number of clusters (K)	Standard deviation of adjusted Rand index		
				K-Means	SBKM	Proposed Algorithm
Data set-1	400	2	2	0.12	0.096	0.045
Data set-2	400	2	2	0.078	0.041	0.034
Data set-3	350	2	3	0.083	0.054	0.046
Iris	150	4	3	0.088	0.041	0.029
Wine	178	13	3	0.082	0.051	0.035

Figure 7: Original Images

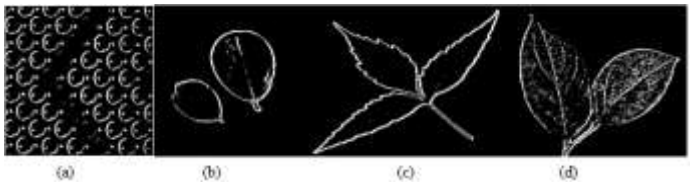

Figure 8: After Edge Detection (Edge pixels as input data points)

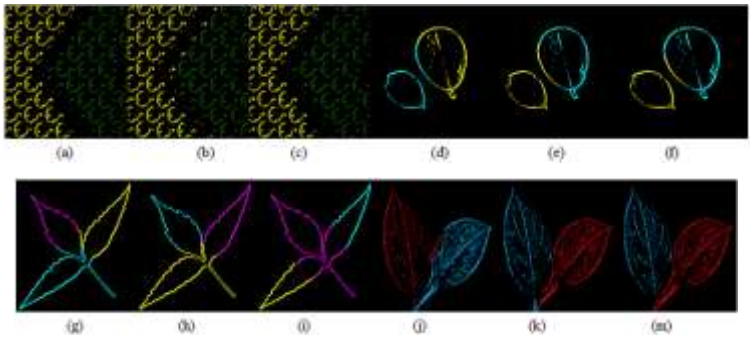

Figure 9:Clustered data: After running the *K*-means algorithm (a, d, g, j), after running the SBKM algorithm (b, e, h, k) and after running the proposed algorithm (c, f, i, m).

REFERENCES

[1] H. Zabrodsky, S. Peleg, and D. Avnir, "Symmetry as a continuous feature," IEEE Transactions on Pattern Analysis and MachineIntelligence, vol. 17, no. 12, pp. 1154–1166, 1995.

[2] M. C. Su and C. H. Chou, "A modified version of the K-means algorithm with a distance based on cluster symmetry," IEEETransactions on Pattern Analysis and Machine Intelligence, 23(6), pp. 674–680, 2001.

[3] C. H. Chou, M. C. Su, and E. Lai, Symmetry as a new measure for cluster validity, 2nd WSEAS Int. Conf. on Scientific Computation and Soft Computing, Crete, Greece, pp. 209.213, 2002.

[4] C.A. Murthy and N. Chowdhury, "In search of optimal clusters using genetic algorithms," Pattern Recognition Letters, vol. 17,no. 8, pp. 825–832, 1996.

[5] G. Garai and B. B. Chaudhuri, "A novel genetic algorithm for automatic clustering," Pattern Recognition Letters, vol. 25, no. 2, pp. 173–187, 2004.

[6] S. Vijendra, K. Ashiwini and S. Laxman, "A fast evolutionary algorithm for automatic evolution of clusters," Information Technology Journal, vol. 11, no. 10, pp. 1409–1417, 2012.

[7] J. Handl and J. Knowles, "An evolutionary approach to multi-objective clustering," IEEE Transactions on Evolutionary Computation, vol. 11, no. 1, pp. 56–76, 2007.

[8] S. Saha and S. Bandyopadhyay, "A symmetry based multiobjective clustering technique for automatic evolution of clusters," Pattern Recognition, vol. 43, no. 3, pp. 738–751, 2010.

[9] W. Pedrycz, Granular Computing: An Emerging Paradigm, vol. 70 of Studies in Fuzziness and Soft Computing, Springer, 2001.

[10] S. Bandyopadhyay, S. Saha, "GAPS: A Clustering Method Using a New Point Symmetry Based Distance Measure", Pattern Recognition, vol. 40, pp. 3430-3451, 2007.

[11] S. Saha, S. Bandyopadhyay, and C. T. Singh, "A new line symmetry distance based pattern classifier," in Proceedings of International Joint Conference on Neural Network, pp. 1425-1432, 2008.

[12] S. Saha and S. Bandyopadhyay, "A new line symmetry distance and its application in data clustering," Journal of Computer Science and Technology, Vol. 24, pp. 544-556, 2009.

[13] S. Saha and S. Bandyopadhyay, "On principle axis based line symmetry clustering techniques," Memetic Computing, Vol. 3, pp. 129-144, 2011.

[14] S. Saha and U. Maulik, "A new line symmetry distance based automatic clustering technique: application to image segmentation," International Journal of Imaging Systems and Technology, 21(1), pp. 86–100, 2011.

[15] R. C. Gonzalez and R. E. Woods, Digital Image Processing, Prentice Hall, Englewood Cliffs, NJ, USA, 2nd edition, 2002.

[16] L. Hubert and P. Arabie, "Comparing partitions," Journal of Classification, vol. 2(1), pp. 193–218, 1985.

[17] D. L. Davies and D.W. Bouldin, "A cluster separation measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 1, no. 2, pp. 224–227, 1979.