Spatial Data Mining with the Application of Spectral Clustering: A Trend Detection Approach

Arvind Sharma
RJIT, Tekanpur
CSE and IT Dept.

R. K. Gupta
MITS, GwaliorCSE and IT Dept.

ABSTRACT

Spectral clustering in spatial data mining plays a very important and innovative role due to its capacity of handling of large size of data ,effective application of linear algebra to solve graphical representation and problems, and application of very low cost of clustering algorithms like k-nearest or ϵ neighbourhood graph. Most of the research in this area is focused on efficient query processing for static or dynamic data. This paper extends the current spatial data mining algorithms to efficient mode of spectral clustering algorithms with the application of Laplacians graph properties and present new approach of spatial data mining methods. These algorithms and methods are used to scratch new knowledge from huge data sets having property of graphs. Obtained results of spectral clustering shows various aspects of spatial data mining and their applications.Spatial database systems contains various spatial objects representing natural objects like mountain or river ,infrastructure like railroad, location, highways with spatial and as well as non spatial attributes. This paper reveals very important and uncovered aspects of spectral clustering.

Keywords

Spectral clustering, Graph Laplacian, spatial data mining, spatial data base systems.

1. INTRODUCTION

Each object has its spatial attributes,location and other non-spatialattributes.These objects and their attribute are applicable in many areas with the help of different clustering approaches.This paper is designed as a better application and exploration of spectral clustering for spatial data mining.Spectral clustering, as its name implies, makes use of the spectrum (or eigenvalues) of the similarity matrix of the data. It examines the *connectedness* of the data, whereas other clustering algorithms such as k-means use the *compactness* to assign clusters.Basically spectral clustering is a large family of combination of methods and it plays active role in research of machine learning, data mining due to its universality, efficiency and supporting result of practical output.

Spectral clustering [3] is becoming more popular due to its vast and useful applications like data analysis, speech separation,video indexing, characterrecognition, image processing and image segmentation etc.

Spectral clustering makes spatial data mining a very important application in following areas-

i. Security deployment: It is based on satellite imaging. The movement of armed forces with vehicles can traced and patterns of movement are detected.
ii. Cellular mobile phone services: Pattern of time, frequency, length, location may be mined. It is used for better policy and planning.
iii. Location awareness and emergency reply: For common objects and geographic similarities we apply clustering technique and spatial clustering is more effective for this application. The spatial location of a user is responsible for location awareness and their profile.
iv. Graph and map based applications. Spectral clustering is also very useful for image segmentation.

In an ideal similarity matrix, the entries between intra-cluster data points are assigned 1, while the entries between inter-cluster data points are assigned 0, we call this the ideal matrix. The use of such an ideal matrix will enable the spectral clustering algorithm to find the exact real clusters.

Spatial data mining is data driven but it is also human centred with full control of users. Selection, integration, transformation, cleaning and application of some special strategies with algorithms makes it applicable and meaningful in different areas of human and scientific life.

SPDM is not easy job, it needs a deep understanding and caution for selection of data and methods to analyse it with certain conditions like measurement, uncertainty, restricted and bias sampling,changeable data units and confidentiality [1].So, it is very important to select SPDM methods and understand them with desired goals and results.As it is mentioned in previous paragraph that selection and analysis method are the planning issues in spatial data mining algorithms. In spatial data mining the clustering is a key concept or stage, during this process and it is mainly used for identification of similarity of objects and patterns which are directly applicable in various social and scientific fields. Some of the common clustering methods are [6] –

i. Density based methods of clustering
ii. Hierarchical based methods of clustering
iii. Partitioning methods of clustering
iv. Grid based methods clustering
v. Constraint based methods of clustering etc.
vi. Spectral clustering based methods

Spectral clustering (vi) is our matter of discussion in this paper.

Spectral clustering [1][2] is better than partitioning methods which is also based on the same concept (k-means etc.). Spectral clustering is especially applicable on moving objects and trajectories, spatially connected social network, spatial information in web based documents, geographic and graphical representation of multidimensional objects etc.

Graph based clustering is key concept of spectral clustering.major strengths of spectral clustering are:

i. Makes no assumptions on the shapes of clusters, can handle intertwined spirals etc.

ii. Provide facility of iterative process to find local minima and multiple restarts.

A sample output is shown in figure 1 which is based on spectral clustering (Laplacian methods).

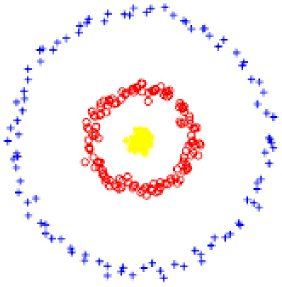


Figure 1:A sample of result of spectral clustering algorithms.

For Spectral clustering the data may be represented in two forms:

1. In table form as

	V1	V2	V3
X	1	3	4
Y	2	1	2

and 2. in graph format as G=(V,E).

Organisation of this research paper is as follows: Section 2 contains basic concept of different clustering methods with graph theory and Laplacian graph which is used in this kind of clustering. section 3, shows existing algorithms and their working performance with positive and their less strengths. Section 4 depicts details of proposed new algorithm of spectral clustering. Section 5 gives an idea about results and discussion. Last section gives detail of future scope and further enhancement.

2. RELATED WORK AND BASIC CONCEPT

In this section we will study the current and previous research work in spatial data mining and knowledge discovery [5]. As we have discussed that clustering plays a very and key role in understanding and application of spatial data in real applications [6]. So we will focus on meaning and methods of clustering in spatial data sets. Recent work in the data base community includes density based methods, hierarchical methods, partition based methods, grid based methods, Spectral clustering methods and constraint based methods[5][6].A brief idea of each and every method is given here with their positive and limited aspects.

2.1Density Based methods: This kind of methods consider clusters as dense region of objects that are different from lower dense regions in the data space. Density based regions are more appropriate applicable in arbitrary shaped clusters but selection of attributes and selection of clusters with algorithms is more complex [6]. It has the feature to merge two clusters that are sufficiently close to each other.

Density Biased sampling, DBSCAN (Density Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure), DENCLUE(Density CLUstEring) etc. are example of this method.

2.2 Hierarchical Based Methods: Hierarchical based methods put the data in a tree like structure. These clusters are classified into agglomerative and divisive hierarchical clustering, depending on whether the decomposition is formed in a bottom up or top down manner.

BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies), CURE (Clustering Using Representatives), CHAMELEON, ORCLUS (arbitrary Oriented projected CLUStergeneration) are the basic methods of this category.

This (Hierarchical methods) can also recognize arbitrary shaped clusters, handles outliers or noise excluding to some special conditions but this method does not work well for special characteristics of individual clusters and it is also time consuming for high dimensional data.

2.3 Partitioning methods: This method divides n objects, which we want to cluster, into k-partitions, where each partition represents a cluster and k is a given parameter. Such algorithms form the clusters to optimize an objective criterion similarity function such as distance as a major parameter.

Partitioning methods cover following five common algorithms as k-means, k-midoids, CLARANS (Clustering Large Applications based up on RANdomized Search).

Although partitioning methods are better in generation of clustering results by using k-mean, k-midoids are easier to implement but selection of n is random so no guarantee of quality of clustering and desired clusters are required in advance which is not more realistic. To handle of outliers of is also a big problem for such kind of methods. A major drawback of this method is to that it is not applicable for large databases.

2.4 Constrained based methods: In our previous paragraphs, we studied that there are so many algorithms and methods to implement and applications of clustering in real life and various social purpose. Unfortunately, most of the algorithms are not able to understand or specify real life constraints such as physical obstacles. So it was realised that there should be a method that can handle the concept of clustering in presence of physical obstacles. These are application specific and used as special cases.

COD-CLARANS(Clustering with obstructed Distance based on CLARANS) is the first clustering algorithm that solve a problem which is known as the problem of clustering with obstacles entities(COE).

This method i.e. constraint based clustering is not well suitable due to NP hard nature of the problems and no guarantee of accuracy of results when number of points are very large that is N.To handle outliers is also a big problem with such kind of methods.

2.5 Spectral clustering: Spectral clustering is a modern type of clustering method and being used as a new approach of clustering. For graph and Laplacians basedapplication it is mainly used with standard concept of mathematics and algebra. When constructing similarity graphs the goal is to model the local neighborhood relationships between the data points which is entirely different from K-means and other methods [6]. The main tool to understand to spectral clustering is Laplacians graph matrices. Very important and popular reason of being successful of spectral clustering is to no consideration or assumption on the basis of clusters forms. Spectral clustering can solve very simple problems like interwined spirals and it is used for large data sets if points are given in the form of sparse. Spectral clustering is used as black box testing method which is the key concept of various clustering and scientific methods.

History: When we go back about spectral clustering and found that it was discussed by Donath and Hoffman(1973). Their concept was to construct graph partitions and it was very good example of eigenvectors and adjacency matrix[5].Similarly ,Fielder proposed that bi – partitions of a graph are also well connected with the next eigenvector of the graph Laplacian and these eigenvectors are applicable to partition a graph.Since then, spectral clustering has become the main research area and improved many times to discover new results. A detailed literature over the history of spectral clustering is given by Spielman and Teng(1996).Some of the positive aspects and why this method is so popular,some points are discussed here:

It was used as machine learning process by the works of shi and malik(2000).Co-clustering problem,distributed environment applications are also extended by the concept of spectral clustering. A very marvellous fact about spectral clustering is that it does not make any assumptions on the type and nature of the clusters only.It has tested and verified that spectral clustering is successfully used over large and huge amount of data sets even similarity graph is sparse and not connected well.It does not suffer from local minima or re initialisation of whole program again and again.Results are more accurate and efficient if we use this clustering with better selection of parameters and care.

All remaining sections demonstrate a complete detail of working of clustering with the help of spectral clustering.

Basic concept behind this clustering method about graph features are discussed here, which arebackbone of understanding the term spectral clustering.

2.5.1 Similarity graphs

To represent similarities between data points, a common and good way of representing the data is in the form of similarity graphs as G=(V,E), where V=vertex and E=Edge[2].

Suppose each vertex v_i represents a data point x_i and if there is a similarity between two points x_i and x_j with a weight with the condition that this weight is positive and greater than with a certain predefined value(threshold)and we represent this similarity as s_{ij}.

The concept of clustering is here to find a partition of the graph as various edges for different groups have very low weights (i.e. points in different clusters are dissimilar from each other) and the edges within a graph have high weights (i.e. points in the same cluster are similar to each other).This concept is very important and plays a key role in organization and application of the term spectral clustering.

Various graph notations are: G(V,E)
V= $(v_1,v_2,\ldots\ldots\ldots\ldots\ldots v_n)$set of vertices
Weighted graph with w_{ij} weight where $w_{ij}\geq 0$.
Weighted adjacency matrix of the graph is - $W(w_{ij})_{ij=1\ldots.n.}$
$w_{ij}=0$ for not connected graph
$w_{ij=}w_{ji}$ for undirected graph.

The degree of vertex v_i є V is defined as
$d_i=\sum_{j=1}^{n} w_{ij}$

D – Degree matrix is diagonal matrix with the degrees d1, d2,…………,dn.

Subset of vertices A ⊂ V and its complement is represented as V \ A by Á.

Indicator vector¶A=$(f_1,\ldots.,f_n)$’ є □$\mathbb{R}^n$ as the vector with entries f_i = 1 if v_iє A and f_i=0 otherwise.
Set of indices {i |v_i є A }, may be represented as for convenience iє A.

If we have two disjoints sets as A,B⊂ V then we define
$W(A,B) := \sum_{i\epsilon\ A,j\epsilon} Wij$

Size of a subset (A⊂V)is measured as
i. |A|:= The number of vertices in A
ii. $Vol(A) := \sum_{i\epsilon A} di$

If there are no connections between A and its complement A and if a subset of A is connected, then it is called connected component. Partition of a graph for non empty sets we have to relations as $A_i \cap A_j = \phi$ and $A_1 \cup \ldots.. \cup A_k = V$.

2.5.2 Basics of Spectral clustering

When we think about spectral clustering for similarity graphs the goal is to determine and find out the local neighbourhood relationship between the data points[8][9].As already have been discussed that data set may be considered as a sequence of data points (x_1, x_2…………………,x_n) and construct pairwise similarities(S_{ij}) or pair distance(d_{ij}) into a graph. Following methods are mainly used for this purpose-

i. The є- neighbourhood graph –It is based on the condition where all points are connected pairwise whose distances are smaller than specially it is used in unweighted graphs. For a given non-negative value є,the є-neighbourhood of an object Oi denoted by Nє(Oi)is defined by Nє(Oi)={ Ojє D | d(Oi,Oj) <= є }
ii. K-nearest neighbour graphs – This concept is also used in hierarchical based clustering methods such as CHAMLEON().Here ,each edge of the graph is weighted to indicate the degree of the similarity between the pair of the data items that are connected by that edge i.e.on edge will weight more if two data objects are similar to each other.The purpose of this method in spectral clustering is to connect vertex vi with vertex vj if vj is among the k-nearest neighbours of vi.If graph is directed then first ignore their directions or mutually k-nearest neighbour property.in both cases,we measure with weight the similarity of endpoints.
iii. The fully connected graph- In this type of graph all pairs are connected by positive similarity calculate weight of each pair as Sij.For the representation of local neighbourhood a similarity function should be used otherwise it will not work properly. A function that is called Gaussian similarity function is define for this purpose. Definition of Gaussian function is given as

$$s(x_i,x_j)=\exp(-\|x_i-x_j\|)^2/2\,\sigma^2\ .$$

Where σ controls the width of the neighbourhoods.This parameter is similar to parameter ϵ as discussed in method I as above.

All types of graph are usable and applicable in spectral clustering.

The main matrices or tools for spectral clustering are graph Laplacian matrices.In this paperfirst we have explained different forms of graph Laplacian and how to use them for spectral clustering .After study of Laplacian based clustering methods (symmetrical or asymmetrical)we will design a new approach of spectral clustering which will be based on Probabilistic sampling of data and then development of new algorithm which will produce better results in terms of space and time.

A little bit background of Laplacian graph is described here for complete understanding of spectral clustering. Here it is assumed that graph G is an undirected graph with weight matrix W[10].According to need of operation and nature of data we will determine eigenvectors of a matrix and necessary eigenvalues. Normally eigenvalues are arranged in increasing order.

Mainly graph is used in un-normalized form of graph Laplacian as

L=D-W
Here L is calculated as f'Lf= f'Df-f'Wf
And L has following properties:

i. f'Lf= 1/2 $\sum_{i,j=1}^{n} w_{ij}\big(f_i - f_j\big)$2. For every vector f $\in \mathbb{R}^n$
ii. L is positive with symmetric property.
iii. L's smallesteigen value is 0 and corresponding eigenvector is the constant value.
iv. Eigenvalues are arranged as 0=λ1≤λ2≤λ3……….≤λn.

This property plays a very important role to describe spectral clustering.Here it should be noted that self-edges in a graph not change the corresponding graph Laplacian.

Another property of graph is based on number of connected components k and the spectrum of L.

It states that:
If G is undirected graph and containing positive values of weights then the multiplicity k of the eigenvalues 0 of L equals the number of connected components A_1,……,$A_K$in the graph.

According to this definition L takes the form

$$L= \begin{pmatrix} L_1 & \cdots & \\ \vdots & L_2 & \vdots \\ & \cdots & L_k \end{pmatrix}$$

Another point should be noted here that each block of L is a proper graph Laplacian.Thus, as a whole the matrix L has many eigenvalues 0 for their connected components k.

Now consider the properties of normalized graph Laplacians – The normalized form of graph Laplacian is based on two terms as symmetric and random walk. Both matrices are defined as

L_{sym}=:= $D^{-1/2}$ $LD^{-1/2}$=I-$D^{-1/2}$W $D^{-1/2}$
L_{rw}:= D^{-1} L=I - D^{-1}W.
Properties of L_{sym}andL_{rw} are as follows :
i. If f$\in\mathbb{R}^n$then

f'L_{sym}f=1/2 $\sum_{i,j=1}^{n} W_{ij}\left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}}\right)^2$

ii. λis an eigenvalue of $L_{rw}$with eigenvector u if and only if λ is an eigenvalue of $L_{sym}$with eigenvector w =$D^{1/2}$ u.
iii. λ is an eigenvalue of $L_{rw}$with eigenvector u if and only if λ and u solve the generalized eigen problem Lu=λD_u.
iv. Here it is also true 0=λ_1≤λ_2≤λ_3……….≤λ_n as above.

3. STUDY OF EXISTING SPECTRAL CLUSTERING ALGORITHMS:

Some notations of graph are already discussed in section 2.Spatial clustering[13] has special importance if our data has large size and arbitrary shape and size[1].Here we consider that each data is a collection of different points "n" as x1,…..xn.Pairwise similarity function as s_{ij}=s(x_i,x_j)and its corresponding similarity matrix as S=$(s_{ij})_{I,j=1….n.}$

Basically there are two forms of spectral clustering for similar graph problems. First is designed especially for un-normalized Laplacian graph[jia et al,2011] and second is used for normalized Laplacian graph.

Un-normalized spectral clustering:
Input :A similarity matrix S∈ℝn*n, Number of clusters to construct=k

Steps:

1. Apply any one method which was described in section 2.2 to construct a similarity graph.Define a weighted adjacency matrix.
2. Determine the L as unnormalized Laplacian matrix.
3. Compute k eigen vectors of L as u1,u2,…………..,uk.
4. Suppose U∈R^{n*n} which contains vectors u1,u2,…………..,uk as coloms.
5. Initialise i= 1 to n ,and let us consider yi∈Rk is a vector which is corresponding to the ith row of U.
6. With the application of k-means algorithm cluster the points into clusters C_1,C_2,……..C_n.

Output: Clusters A_1,……..,A_k with A_i={j|y_j∈C_i}.

Normalized form of spectral clustering: For symmetric algorithm we use both Laplacian graph as L_{sym} or L_{rw} with Eigenvectors.Here all steps are same excepting to steps 4 and 5.

Algorithm:
Input: Similarity matrix have to be construct as S∈Rn×n and k number of clusters has to be created[13].

Steps:

1. Design and construct a similarity graph by one of the ways described in section 2.2
2. Determine and calculate the normalized Laplacian L(L_{sym} or L_{rw}).
3. calculate and identify first k-eigenvectors u_1,u_2,……..u_k of L.
4. Suppose U∈Rn×k be the matrix containing the vectors u_1,u_2,……..u_k as columns.
5. Construct a matrix T∈$\mathbb{R}^{n\times k}$ from U by normalizing the rows to norm1, that is sett_{ij}=$u_{ij}/\left(\sum_k u_{ik}^2\right)^{1/2}$.
6. For i= 1 to n ,lety$_i$∈$\mathbb{R}^k$ be the vector processing to the i^{th} row of T.

7. Cluster the point (y_i),i=1,2,…………,nwith the k-means algorithm into clusters C_1,……C_k.

Output: Clusters A_1,……A_k with $A_i=\{j|y_j \in C_i\}$.

Although above mentioned algorithms are benchmark in the research field of spectral clustering but as time passes everything needs improvement in terms of data selection,nature of learning(supervised/unsupervised), space and time efficiency [5].

After thoroughly study we found following limitation in above mentioned algorithms;

1. These traditional algorithms can group only small data sets of input.
2. The final time complexity of Laplacian matrix is O(n3) with application of Eigen decomposition.
3. Only to store similarity matrix, above mentioned algorithms take O(n2) space and this is the reason to increase total time complexity.
4. For large data sets iterations and computations are time consuming process.
5. Accuracy of clusters is not so accurate.

To overcome all these problems, a different and new algorithm is proposed in next section.

4. OUR PROPOSED ALGORITHM FOR BETTER RESULTS (SR-SC)

Before designing new algorithm, it is very important to understand about Eigenvalues and Eigenvectors in matrix theory[2]. A little bit idea is given here and it is assumed that whosoever is working in this field already has detailed knowledge of these terms and their method of calculation.

Let A be a square matrix. Assume λ is an eigenvalue of A. In order to find the associated eigenvectors, we do the following steps:

1. Write down the associated linear system
   $AX = \lambda X$ or $(A-\lambda I_n)X = O$
2. Solve the system.
3. Rewrite the unknown vector *X* as a linear combination of known vectors.

These results of eigenvectors are used to map the data points to a lower dimension and then apply k-means(mostly) algorithm to make groups in the form of clusters[2]. Normalized cut and minimum cut play major roles to complete this clustering procedure. A brief introduction of graph theory is given in section 1 for understanding of the importance and application of these theories and then we get better results of spatial clustering.

To get low dimensional embedded space, It is needed to break $I-D^{-1/2}W\ D^{-1/2\ to}$ determine eigenvalues and eigenvectors as

$(D^{-1/2}W\ D^{-1/2})V=V\Lambda$

Where V is a $n_E X n_E$matrix and formed by eigenvectors and Λ is a diagonal matrix and it is formed by eigenvalues.Our proposed algorithm is Sampling Ratio based Spectral Clustering(SR-SC) as mentioned in heading of section 4.

A brief idea behind this algorithm is given here[14]. First we will divide all data points in two parts,one point is m random sampled and remaining n-m data points are used to further process.A similarity matrix W is calculated where A belongs to sampled data points and B contains remaining points of similarity matrix.Other necessary symbols are defined in the algorithm.

A proposed algorithm about spectral clustering which is based on probability of sampling in incremental form is given here[5][7]:

Algorithm:

Input: A data set $D=\{d_{i,}i=1,….,n\}$
The number of samples m
Desired number of cluster k
Output: k- grouped clusters

Steps:
Step1. Calculate pairwise similarity between given data points of set D.

As $W \in \mathbb{R}^{n\times n}$

And where $w_{ij} = \exp(-||x_i-x_j||^2/2\sigma^2)$ where σ may be improved by certain procedure and conditions.

Step 2. Update each row of similarity matrix for sampling probability as $p \leftarrow\ p/||p||_2$.After it, form the similarity matrix as $A \in \mathbb{R}^{m\times m}$ and $B \leftarrow \mathbb{R}^{m\times(n-m)\cdot}$Where A represents sampled points and B represents remaining points without sampled.

Step 3. Compute the degree of nodes $(\hat{d})$ for matrix A and B by using formula

$$\hat{d}=\bar{W}_l=\begin{bmatrix} ar+br & \\ b_c+B^TA^{-1} & br \end{bmatrix}$$

Where a_r and $b_r \in \mathbb{R}^m$represent the row sum of A and B ,respectively.$b_c \in \mathbb{R}^{n-m}$represent the column sum of B, and l is column vector with all 1 elements.Nowwe will normalize matrix A and B.

Step 4. Calculate matrix Q with the normalized A and B as $Q= A+A^{-1/2}BB^TA^{-1/2}$.

Step 5. Determine Matrix $\cup_Q$ and $\wedge_Q$by decomposing of Q as $Q=\cup_Q\wedge_Q\cup^T{}_Q$

Step 6. Put values of $\cup_Q$and$\wedge_Q$ Then finally calculate the orthogonal eigenvector of $\bar{W}$
AsV=[A] $A^{-1/2}\cup_Q\wedge_Q{}^{-1/2}$

Step 7. Select the Eigen vectors corresponding to the first k-largest eigenvalues V_1,………..,V_k of $\bar{W}$ from to and then compose matrix
V_k:$V_k=[_{Vi},…………..,V_k] \in \mathbb{R}^{n\times k}$.

Step 8. Normalize each row of V^k to a unit vector and obtain matrix Y with each element
$Y_{ij}=V_{ij}/\sqrt{\Sigma^k}\ V_{ij}{}^{2j=1}$

Step 9. Matrix found in step 8 forms a space and this space is used to get clusters by application of standard algorithm of spectral clustering (It may be K-Means algorithm).

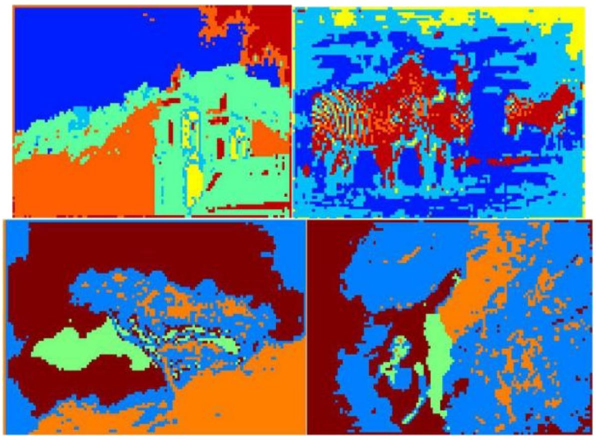
5. RESULTS AND DISCUSSION

Below some of the images are given and their outcomes after application of normalized spectral clustering methods.In section 4, we proposed a novel and improved algorithm with the help of probabilistic sampling concept.

Results and their meanings are shown in Table 1 and Figure2, figure 3 of Laplacian based symmetrical method with random walk.

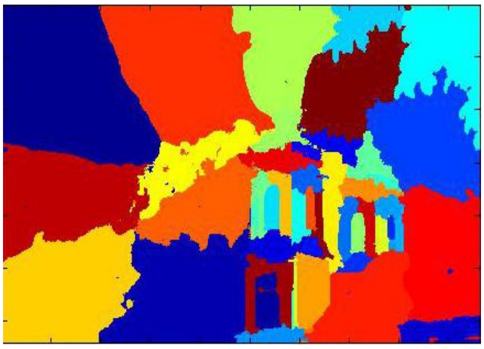
(a) Original image (b) Original image
(c) Original image (d) Original image

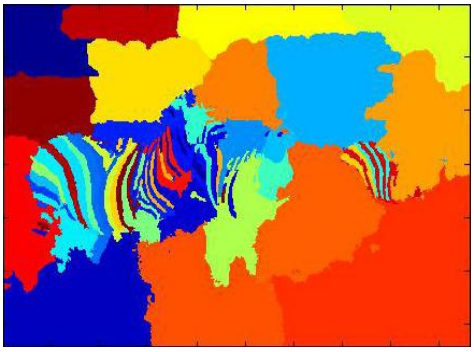
In figures (a) to (d) different images are collected from different sources and the process of normalized clustering algorithms has applied there for getting clustered and segmented form of images. Results of above images are shown below as figure (e) to (h).



(e) Result of (a) (f) Result of (b)
(g) Result of (c) (h) Result of (d)

Figure 2: Original images and their normalized outputs





(i)Multi scale clustering where clusters =45 for fig. (a)
(j)Multi scale clustering where clusters =40for fig. (b)

Figure 3: Results after deciding number of clusters

As shown in figures 2 and 3 that spectral clustering is a major key task to get clusters of graph based data, with and without normalized process. After little bit changes in algorithms we can get desired output in terms of number of clusters for spatial data bases[8][9].

We implement the proposed spectral clustering algorithm with random walk (SCRW) and compare with the traditional spectral clustering (SC), local scale spectral clustering(LSSC) and k-means clustering. We test the performance of these algorithms on some benchmark data sets.

We use three widely used synthetic data sets (FourLine, ThreeCircle, and CirclePoints) to evaluate the proposed algorithm. The steps of the random walk is set to 50. The clustering results are shown in Fig 4 for four line clusters. Different lines are shown by different clusters.These results may be converted into three circle or circle point representation.This is shown for synthetic data or artificial data.

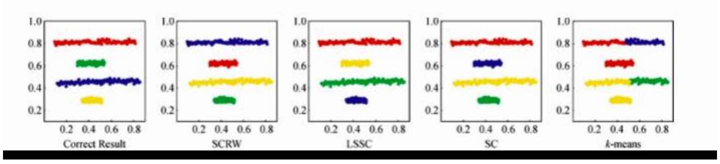


Figure 4: Four-line data sets results are shown for different algorithms.

Results of real data is shown in fig4.We choose digits{1,7}, {8, 9}, {0, 8} and {1, 2, 3, 4} as subsets for experiment separately as shown in table1.InTable1,errorrate is used to evaluate the performance of different clustering methods[4]. Compared to other algorithms, the proposed algorithm has a lower error rate and is more robust for clustering.

In spectral clustering we choose σ from1 to 200,and then use the best one as final result. In LSSC(Local Scale Spectral Clustering), we choose the distance to its 7-th nearest neighbour, as the local scale. While in SCRW(Spectral Clustering based on Random walk), we choose the random walk steps M to be 101. Results of these algorithms are described and shown in table 1 and figure 4.Selection of data sets have been taken from various sources as agriculture,wine, chemical analysis of chemical materials ,TV and RADAR data and many more but we have selected here as shown in table1[11][12].

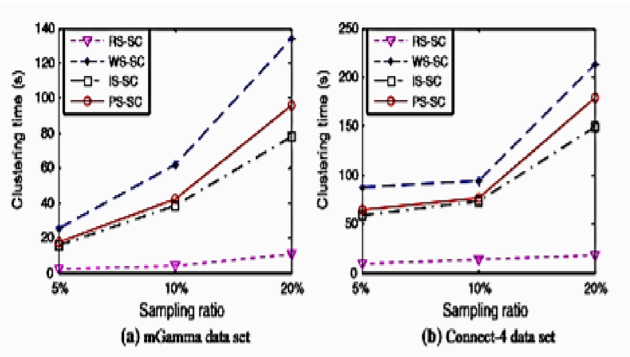
Table-1 : Error rate for different algorithms

Data	SCRW	LSSC	SC	K-Means
Iris	0.093330	0.093330	0.32000	0.11333
Wine	0.275280	0.452510	0.38547	0.46980
{1,7}	0.026760	0.038930	0.01945	0.00973
{0,8}	0.030470	0.190460	0.18667	0.01714
{8,9}	0.023320	0.160350	0.17492	0.07871
{1,2,3,4}	0.038640	0.655790	0.09299	0.10024
Glass	0.457944	0.500000	0.50000	0.514019
Ionosphere	0.210826	0.264957	0.54700	0.293447

6. RESULTS OF OUR DEVELOPED AND PROPOSED ALGORITHM FOR TREND DETECTION AND PREDICTION IN SPATIAL DATABASES

Table-2: Results of different algorithms for different data sets[11],[12] with different sampling ratio.

Sampling Ratio m/n	Data Set	SRSC	WSSC	ISSC	PSSC
5%	ImageSeg Musk PenDigits mGamma Connect-4 USCI Poker Hand	0.028 0.231 1.673 2.098 9.743 33.076 60.094	0.563 2.034 15.063 25.671 87.673 258.765 493.456	0.218 1.581 9.534 15.456 59.456 81.342 220.721	0.237 1.562 8.732 17.453 64.234 176.786 221.093
10%	ImageSeg Musk PenDigits mGamma Connect-4 USCI Poker Hand	0.277 0.642 3.234 4.564 13.231 45.782 73.562	1.432 5.678 38.871 61.341 93.894 371.098 864.468	0.441 2.435 17.654 38.234 73.091 265.782 732.513	0.573 3.439 21.563 42.769 76.902 284.096 577.761
20%	ImageSeg Musk PenDigits mGamma Connect-4 USCI Poker Hand	0.263 0.765 8.456 10.234 18.765 66.870 117.896	4.365 14.234 77.231 134.567 213.641 587.871 1293.547	1.697 6.456 42.321 78.095 149.562 308.670 813.450	1.342 9.521 54.679 95.713 178.459 325.892 876.453



RSSC or SRSC-Sampling Ratio based Spectral Clustering
WSSC-Spectral Clustering on Weighted sampling
ISSC-Spectral clustering based on Incremental Clustering
PSSC-Probability Sampling based Spectral Clustering

Figure 5:Performance graph representation for most popular algorithms and our developed algorithm.

Table -2 and figure 5 depict that large number of sample points take more time for generating clusters for different algorithms and different ratio of sampling .It is shown in figure and table that our algorithm takes less time than other popular algorithms.As shown in above mentioned figure and table that as far as we increase sampling ratio as 5,10 and then 20% then required running time is also increased by a certain factor[7] .For large data sets and sampling points ,time complexity is also increases and that is future scope and challenge to other researchers that how to reduce this time with application of novel and intelligent steps of processing on our algorithm(SR-SC).

7. FUTURE SCOPE AND LIMITATIONS

As mentioned in introduction section that spectral clustering plays very important role in various areas of scientific research like character recognition, videoindexing, image processing and image segmentation but still it requires some advancement in important areas and future prospects. These future prospects are

i. Still application of priory knowledge for solving semi supervised spectral clustering based problems.
ii. Hard divisions of input data with the concept of classical set theory is not perfect for some applications.So, soft division is necessary to avoid ambiguity in results or clusters. So soft division of objects and classes is still needed for future work.

For complex distribution of objects or samples,procedure produces result where accuracy matters and it goes difficult to identify in results.For this purpose, we can suggest kernel methods with the integration of spectral clustering for better performances.

As we studied and have shown different results of clustering algorithms and it is basic fundamental point is notable that performance of the algorithm depends on time and space complexity. So, cubic time complexity and quadratic time space with regard of data size is also a considerable part for getting better results i.e. the complexities ofalgorithms may be optimized.In other words, we can say that application of multiple methods and improved algorithms is always an open challenge for researchers to improve results in more accurate and advanced manner.

8. REFERENCES

[1] Bach, F. and Jordan,M (2004), Learning spectral clustering. pp 305-312,Cambridge, MA:MIT press.

[2] Dhillon,I, Guan,Y.,andKulis,B(2005).A unified view of kernel k-means, spectral clustering, and graph partitioning.

[3] Kempe, D. and Mcsherry, F.(2004).A decentralized algorithm for spectral analysis (PP 561-568), NY, USA; ACM press.

[4] M. Audibert, J-Y.,and Von Luxburg ,U(2007)Graph Laplacian and their convergence on random neighbourhood graphs.JMLR,8,1325-1370.

[5] Xianchao ZHANG at al, Dalian university of technology" An improved spectral clustering algorithm based on random walk". Springer-Verlag Berlin, 2011(PP 268-278)

[6] Arvind Sharma, R K Gupta at al (2016). "Improved DBSCAN", Hindawi publication

[7] Jia H., Ding S. at al. "A Nystrom spectral clustering algorithm based on probability incremental sampling" DOI 10.1007/s00500-016-2160-8.

[8] Liu T., Gu Y., at al "Class specific sparse multiple kernel learning for spectral – spatial hyperspectral image classification" IEEE transactions on Geo Science and Remote sensing-2016.

[9] H. jia. S. Ding.X. Xu. "The latest research progress on spectral clustering" DOI 10.1007/s00521-s013-1439-2, Springer-Verlag,2013.

[10] Ding, C., He, X., Zha,H., et al: A Min-Maxcut algorithm for graph partitioning and data clustering. pp 107-114(2001), California .

[11] Open Geospatial Consortium (OGC) http://www.opengeospatial.com

[12] Office of Policy Development and Research (PD&R) U.S. Department of Housing and Urban Development. www.huduser.gov

[13] L.,Ulrike von. "A Tutorial on Spectral clustering",17(4),2007,Germany.

[14] Hendrickson, B. And Leland, R (1995). An improved Spectral graph partitioning algorithm for mapping parallel computations. SIAM J. On Scientific Computing, 16, 452-469.