iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** I **Month of publication:** January 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue I, January 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Fast and Accurate Spectral Clustering Based KNN-Similarity Graph Analysis

S. Shanmugaprabha[1], R. Sekar[2]
[1]Research Scholar, M.phil, NGM College, Pollachi, India.
[2]Research Guide, Department of IT,NGM College, Pollachi,India.

Abstract— The recent years as an important analytical technique, both due to the prevalence of graph data, and the usefulness of graph structures for exploiting intrinsic data characteristics. However, as graph data grows in scale, it becomes increasingly more challenging to identify clusters. The propose an efficient clustering algorithm for large scale data using spectral methods. Finding clusters in data is a challenging task when the clusters differ widely in shapes, sizes, and densities. The proposed system present a novel spectral algorithm with a similarity measure based on modified nearest neighbor graph. The resulting affinity matrix reflexes the true structure of data. Its eigenvectors, that do not change their sign, are used for clustering data. The algorithm requires only one parameter a number of nearest neighbors, which can be quite easily established. Its performance on both synthetic and real data sets is competitive to other solutions.

Keywords— Spectral Cluster, KNN Graph, Affinity Matrix, Similarity Graph.

I. INTRODUCTION

Clustering has been extensively explored as one of the most fundamental techniques in machine learning and data mining. Various applications, such as image segmentation, gene expression analysis, document analysis, content based image retrieval, image annotation, similarity searches, have witnessed the practical effectiveness of clustering.

Clustering is one of the most widely used techniques for exploratory data analysis, with applications ranging from statistics, computer science, and biology to social sciences or psychology. In virtually every scientific field dealing with empirical data, people attempt to get a first impression on their data by trying to identify groups of “similar behaviour” in their data. In this article we would like to introduce the reader to the family of spectral clustering algorithms. Compared to the “traditional algorithms” such as k-means or single linkage, spectral clustering has many fundamental advantages. Results obtained by spectral clustering often outperform the traditional approaches, spectral clustering is very simple to implement and can be solved efficiently by standard linear algebra methods.

Spectral clustering (SC) has gradually become one of the most important clustering techniques and it shows more capability in partitioning data with more complicated structures compared to traditional clustering approaches. The underlying reason is that spectral clustering puts more efforts on mining the intrinsic data geometric structures. SC has been widely applied and shown their effectiveness in various real-world applications, such as image segmentation. The fundamental idea of spectral clustering is that it predicts cluster labels by exploiting the different similarity graphs of data points. Besides NCut and k-way NCut, a new SC algorithm, i.e., local learning based clustering (LLC), was developed according to the assumption that the cluster label of a data point can be determined by its neighbors, and a kernel regression model was used for label prediction. There are several popular constructions to transform a given set x1…,xn of data points with pairwise similarities sij or pairwise distances dij into a graph. When constructing similarity graphs the goal is to model the local neighborhood relationships between the data points.

The main purpose of spectral clustering algorithm that can simultaneously address both of the above mentioned challenges for a variety of data sets. In propose algorithm the similarity between pairs of points is deduced from their neighborhoods. The use of similarity based on nearest neighbors approach removes, at least partially, problems with cluster varying densities and the unreliability of distance measure. Resulting adjacency matrix reflects true relationships between data points. Apart from only one parameter another advantage of the presented approach is that it incorporates a variety of recent and established ideas in a complete algorithm which is competitive to current solutions.

II. RELATED WORK

On the one hand, the intrinsic correlations among multiple clustering tasks on different yet related data, namely inter task correlation, are inevitably overlooked in traditional single-task clustering approaches, such as k-means and fuzzy c-means. On the other hand, although to some extent the inter task correlations are explored in previous works to enable multitask clustering, they have been consistently confronted with another serious challenge, i.e., no effective mechanism is afforded to deal with the out-of-sample data,

which is especially significant confronted with the current evolution of Web data. A promising way is to learn an explicit mapping function for predicting cluster labels for the out-of-sample data outside the training data. Several approaches have been proposed to provide an additional step to patch the problem, such as Nystrom method. However, such approaches normally separate learning cluster labels and learning mapping functions into two individual steps, thereby ignoring the relations between them. Moreover, while most of the previous works focus more on exploring the inter task relations, the within-task properties are not well considered. Although in traditional spectral clustering, by exploiting data local structures we may improve clustering performance to some extent, the task-specific property of the cluster indicator matrix requires more investigation. For instance, more discriminative information should be embedded in the cluster indicator matrix to make the clustering algorithms more effective and solid. Also, under some circumstances, the over fitting problems may occur and degrade the clustering performance due to the lack of appropriate processing.

They study the spectral properties of an adjacency matrix A and its connection to the data generating distribution P. The authors investigate the case when the distribution P is a mixture of several dense components and each mixing component has enough separation from the others. In such a case A and L are (close to) block-diagonal matrices. Eigenvectors of such block-diagonal matrices keep the same structure. For example, the few top (i.e. corresponding to highest eigen values) eigenvectors of L can be shown to be constant on each cluster, assuming infinite separation between clusters. This property allows distinguishing the clusters by looking for data points corresponding to the same or similar values of the eigenvectors. The existing system develops theoretical results based on a radial similarity function with sufficiently fast tail decay. They prove that each of the top eigenvectors of A corresponds exactly to one of the separable mixture components. The eigenvectors of each component decay quickly to zero at the tail of its distribution if there is a good separation of components. At a given location xi in the high density area of a particular component, which is at the tails of other components, the eigenvectors from all other components should be close to zero.

The existing system attempts to characterize eigenvectors of the laplacian on regular graphs. He suggests that the distribution of eigenvectors, except the first one, follows approximately a Gaussian distribution. There are also proofs that in general, top eigen values have associated eigenvectors which vary little between adjacent vertices. The two facts confirm the assumption that each cluster is reflected by at least one eigenvector with large components associated with the cluster vertices and almost zero values in the other case. Concept incorporated in the algorithm is so-called modularity, i.e. a quality function introduced by Newman for assessing a graph cut. According to its inventor a good division of a graph into partitions is not merely one in which there are few edges between groups; it is one in which there are fewer than expected edges between groups. The modularity Q is, up to a multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent graph with edges placed at random, or in functional form.

The general spectral clustering method was first shown to work on data represented in feature space. As we are mainly interested in graph data, we need one more step to construct an adjacency matrix which takes O(n2p) time where n and p represent number of data points and features respectively. Calculating the eigen decomposition of the corresponding laplacian matrix is the real computational bottleneck, requiring O(n3) time in the worst case. Therefore, applying spectral clustering for large scale data becomes impossible for many applications. In recent years, many works have been devoted to accelerating the spectral clustering algorithm.

Similar to this idea, , all data points are collapsed into centroids through k-means or random projection trees so that eigen-decomposition only needs to be applied on the centroids. The existing uses random projection in order to reduce data dimensionality. Random sampling has also been applied to reduce the size of data points within the eigen-decomposition step. In, landmark points are first selected among all the data points to serve as a codebook. After encoding all data points based on this codebook, acceleration can be achieved using the new representation. The resistance distance embedding, this employs a similar idea to spectral clustering and exhibits comparable clustering capability.

III.PROPOSED ALGORITHM

A. Similarity Graph

Given a set of data points x_1 ..... and some notion of similarity $S_{ij} \geq$ between all pairs of data points and , the intuitive goal of clustering is to divide the data points into several groups such that points in the same group are similar and points in different groups are dissimilar to each other. If we do not have more information than similarities between data points, a nice way of representing the data is in form of the similarity graph $G = (V, I$. Each vertex vi in this graph represents a data point xi. Two

vertices are connected if the similarity between the corresponding data points and , is positive or larger than a certain threshold, and the edge is weighted by . The problem of clustering can now be reformulated using the similarity graph: we want to find a partition of the graph such that the edges between different groups have very low weights (which means that points in different clusters are dissimilar from each other) and the edges within a group have high weights which means that points within the same cluster are similar to each other.

B. Graph Notation

Let be an undirected graph with vertex set . In the following we assume that the graph G is weighted, that is each edge between two vertices vi and vj carries a non-negative weight . The weighted adjacency matrix of the graph is the matrix . If this means that the vertices and are not connected by an edge. As G is undirected we require . The degree of a vertex is defined as

$$d_i = \sum_{j=1}^{n} w_{ij} \qquad \text{eq. (1)}$$

Note that, in fact, this sum only runs over all vertices adjacent to v_i, as for all other vertices vj the weight is 0. The degree matrix D is defined as the diagonal matrix with the degrees on the diagonal. Given a subset of vertices , we denote its complement by . We define the indicator vector as the vector with entries if and if otherwise. For convenience we introduce the shorthand notation for the set of indices , in particular when dealing with a sum like . For two not necessarily disjoint sets define

$$W(A, B) := \sum_{i \in A, j \in B} w_{ij} \qquad \text{eq. (2)}$$

Consider two different ways of measuring the "size" of a subset $A \subset V$:

$$|A| := \text{the number of vertices in } A$$

$$Vol(A) := \sum_{i \in A} d_i \qquad eq. (3)$$

Intuitively, |A| measures the size of A by its number of vertices, while vol(A) measures the size of A by summing over the weights of all edges attached to vertices in A. A subset of a graph is connected if any two vertices in A can be joined by a path such that all intermediate points also lie in A. A subset A is called a connected component if it is connected and if there are no connections between vertices in A and . The nonempty sets $A_1, \ldots, A_k$ form a partition of the graph if $A_i \cap A_j = \emptyset$ and $A_1 \cup \ldots \cup A_k =$

C. K-nearest neighbor graphs

The goal is to connect vertex vi with vertex v_j if v_j is among the k-nearest neighbors of vi. However, this definition leads to a directed graph, as the neighborhood relationship is not symmetric. There are two ways of making this graph undirected. The first way is to simply ignore the directions of the edges, that is we connect v_i and v_j with an undirected edge if vi is among the k-nearest neighbors of v_j or if v_j is among the k-nearest neighbors of v_i. The resulting graph is what is usually called the k-nearest neighbor graph. The second choice is to connect vertices v_i and v_j if both v_i is among the k-nearest neighbors of v_j and v_j is among the k-nearest neighbors of vi. The resulting graph is called the mutual k-nearest neighbor graph. In both cases, after connecting the appropriate vertices weight the edges by the similarity of their endpoints.

IV. GRAPH LAPLACIAN

The main tools for spectral clustering are graph Laplacian matrices. There exists a whole field dedicated to the study of those matrices, called spectral graph theory. In define different graph Laplacians and point out their most important properties. We will carefully distinguish between different variants of graph Laplacians. Note that in the literature there is no unique convention which matrix exactly is called "graph Laplacian". Usually, every author just calls "his" matrix the graph Laplacian. Hence, a lot of care is needed when reading literature on graph Laplacians. In the following we always assume that G is an undirected, weighted graph with weight matrix W, where . When using eigenvectors of a matrix, we will not necessarily assume that they are

normalized. By "the first k eigenvectors" we refer to the eigenvectors corresponding to the k smallest eigen values.

A. Unnormalized graph Laplacian

The unnormalized graph Laplacian matrix is defined as

$$L = D - W \quad eq.(4)$$

An overview over many of its properties. The following proposition summarizes the most important facts needed for spectral clustering. the un-normalized graph Laplacian does not depend on the diagonal elements of the adjacency matrix W. Each adjacency matrix which coincides with W on all off-diagonal positions leads to the same un-normalized graph Laplacian L. In particular, self-edges in a graph do not change the corresponding graph Laplacian.

B. Normalized graph Laplacians

There are two matrices which are called normalized graph Laplacians in the literature. Both matrices are closely related to each other and are defined as

$$L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \quad eq.(5)$$

$$L_{rw} = D^{-1} L = I - D^{-1} W \quad eq.(6)$$

The first matrix by Lsym as it is a symmetric matrix, and the second one by Lrw as it is closely related to a random walk. In the following we summarize several properties of Lsym and Lrw. The standard reference for normalized graph Laplacians.

V. SPECTRAL CLUSTERING ALGORITHMS

They like to state the most common spectral clustering algorithms. For references and the history of spectral clustering we refer to Section 9. We assume that our data consists of n "points" x1, . . . , xn which can be arbitrary objects. We measure their pairwise similarities sij = s(xi, xj) by some similarity function which is symmetric and non-negative, and we denote the corresponding similarity matrix by S = (sij)i,j=1...n.

Steps: Unnormalized spectral clustering

Step 1: Input: Similarity matrix $S \in \mathbb{R}^n$, number k of clusters to construct.

Step 2: Construct a similarity graph by one . Let W be its weighted adjacency matrix.

Step 3: Compute the unnormalized Laplacian L.

Step 4: Compute the first k eigenvectors $u_1 \ldots \ldots$ of L.

Step 5: Let $U \in \mathbb{R}^n$ be the matrix containing the vectors $u_1 \ldots \ldots$ as columns.

Step 6: For $i = 1, .., $ let $y_i \in \mathbb{R}$ be the vector corresponding to the i-th row of U.

Step 7: Cluster the points $(y_i)_{i=1,}$ in with the k-means algorithm into clusters $C_1, \ldots, C_k$.

Step 8: Output: Clusters $A_1, \ldots, A_k$ with $A_i = \{j | y_i \in\}$.

There are two different versions of normalized spectral clustering, depending which of the normalized graph Laplacian is used.

VI. EXPERIMENTAL RESULTS

Here have compared the performance of the KNN similarity algorithm (implemented in MATLAB) to three other methods: the Ng et al. algorithm, the Fischer et al. algorithm, and the multitask algorithm.

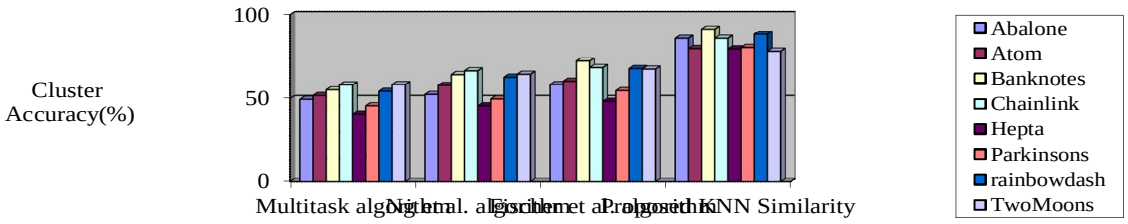


Fig.1. Compare Existing with proposed cluster accuracy

The cluster accuracy showed in Fig. 1 show compare existing algorithm with proposed algorithm. Finaly proposed algorithm give

Volume 4 Issue I, January 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

high accuracy compare with existing.

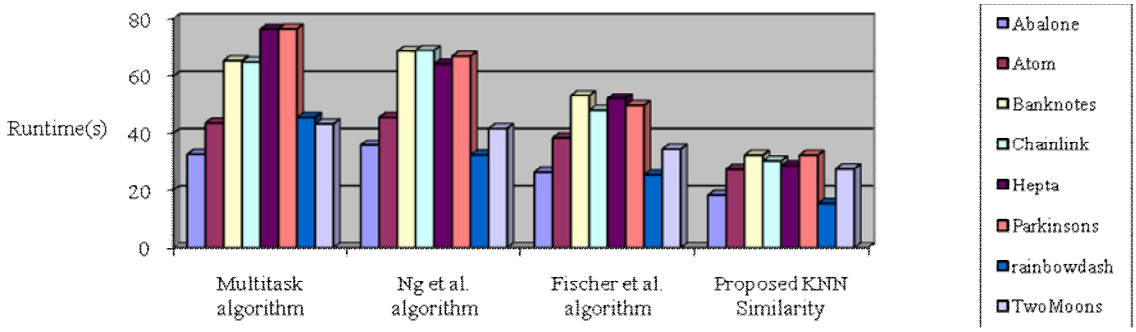


Fig.2. Compare Existing with proposed cluster runtime

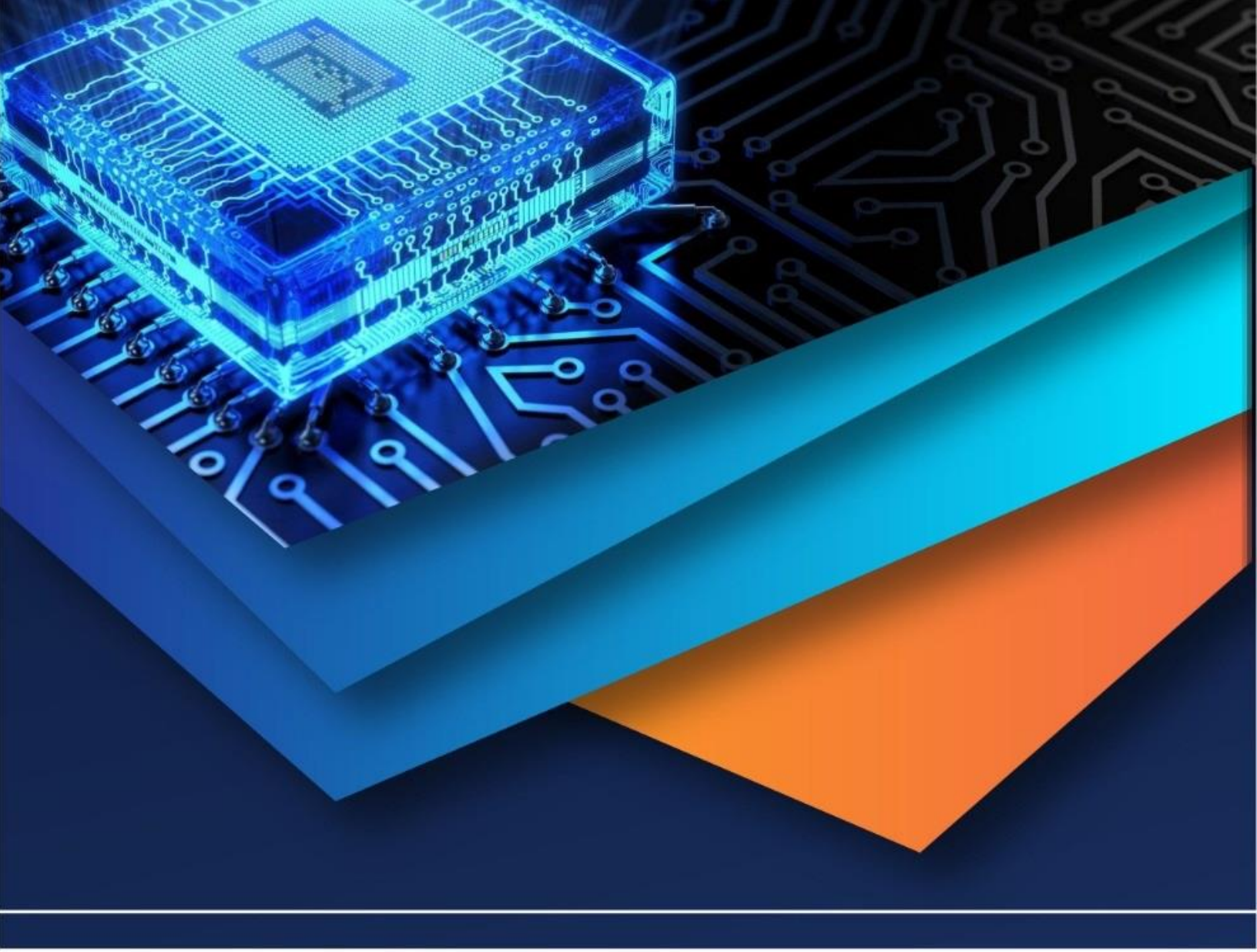
The cluster runtime showed in Fig. 1 show compare existing algorithm with proposed algorithm. Finally proposed algorithm give low time compare with existing.

VII. CONCLUSION AND FUTURE WORK

Spectral clustering can be implemented efficiently even for large data sets, as long as we make sure that the similarity graph is sparse. Once the similarity graph is chosen, we just have to solve a linear problem, and there are no issues of getting stuck in local minima or restarting the algorithm for several times with different initializations. Automatically detects the correct clusters in any given data set. But it can be considered as a powerful tool which can produce good results if applied with care. The proposed method can considerably decrease the necessary runtime while posting a tolerably small loss in accuracy. The future work observations show that our algorithm is a good candidate to apply it to image segmentation, that will be our next task.

REFERENCES

[1] Cvetkovic D.: Signless Laplacians and line graphs. Bull. Acad. Serbe Sci. Arts, Cl. Sci. Math. Natur., Sci. Math. 131, No. 30, pp. 85-92 (2005).
[2] Deepak, V., and Meila, M.: Comparison of Spectral Clustering Methods. UW TR CSE-03-05-01 (2003).
[3] Elon, Y.: Eigenvectors of the discrete Laplacian on regular graphs a statistical approach. J. Phys. A: Math. Theor.41 (2008).
[4] Fischer, I., and Poland, J.: Amplifying the Block Matrix Structure for Spectral Clustering. Technical Report No. IDSIA-03-05, Telecommunications Lab (2005).
[5] Jain, A. Murty, M., and Flynn, P. Data clustering: A review. ACM Computing Surveys, 31, pp. 264-323 (1999).
[6] Jain, A.: Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31, pp. 651-666 (2010).
[7] MacQueen, L.: Some methods for classification and analysis of multivariate observations. In: LeCam, L. and Neyman, J. (eds.) 5th Berkeley Symposium on Mathematical Statistics and Probabilitz, vol. 1, pp. 281-297. University of California Press, Berkeley (1967).FLEXChip Signal Processor (MC68175/D), Motorola, 1996.
[8] Maier, M., Hein, M., and von Luxburg, U.: Cluster identification in nearest-neighbor graphs. In: Proc. of the 18th International Conference on Algorithmic Learning Theory, ALT'07, Springer, Berlin, Germany, pp. 196-210 (2007).
[9] Meila, M., and Shi, J.: A random walks view of spectral segmentation. In: Proc. of 10th International Workshop on Arti¯cial Intelligence and Statistics (AISTATS), pp. 8-11 (2001).
[10] Newman, M.E.J.: Detecting community structure in networks. European Physics. J. B 38, pp. 321-330 (2004).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY