International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-04, Issue-09, Dec 2018

A Novel K-means Spectral Graph Partitioning Based Data Clustering for Detection of Risk Factor in Diabetic Patients

Abinesh S[1], G. Prabakaran[2] and R. Arun Kumar[3]

[1]Research Scholar, Dept. of Computer Science & Engg, Annamalai University, Chidambaram

wap.abineshs2018@gmail.com

[2,3]Assistant Professor, Dept. of Computer Science & Engg, Annamalai University, Chidambaram

prabakaran.g.5492@annamalaiuniversity.ac.in[2], arunkumar.r.8130@annamalaiuniversity.ac.in[3]

Abstract: - Clustering is a technique used widely in data mining and data analysis applications using a set of data points. It is used for separate data points into groups, based on the similarity many numerical methods which are currently available lagging in its performance due to its probability distribution. Clustering in medical applications plays a vital role to detect the abnormalities in patients. The proposed research work addresses the risk factor in diabetic patients using clustering algorithms. Spectral clustering is used in data set partition to create clusters in un-weighted and vertex weighted approaches. The use of vectors in partitioning the vertex is near to zero while compared for larger coefficients of graph values. Spectral graph clustering incorporates the vectors and their vertex weight for each given graph is suitable for data analysis applications. Combining k-means algorithm with spectral graph partitioning in the proposed research work for detecting the abnormalities and provides improved time efficiency for large-scale datasets present in medical field compared to existing conventional clustering algorithms.

Keywords: - Spectral clustering, cluster analysis, unsupervised clustering, K-means

I. INTRODUCTION

Data clustering is widely used in various fields depending upon the applications such as segmentation in image processing, graph partitioning in parallel computing. The rapid development of information technology over past decades multiplying the complexity of processing large data and the response of handling such resource needs a better psychological measurement. A diabetic is a metabolic disorder and complicated diseases which is spread worldwide. International diabetes federation provides health threat in worldwide that 380 million people are suffering from diabetes. In India out of 380 million 65 million people suffered in India alone and in that 40 million cases are undiagnosed. Various image processing models are used in detection of medications in patients. Reliable methods are available for assessment of medical data which assists the professionals and reduce the examination time.

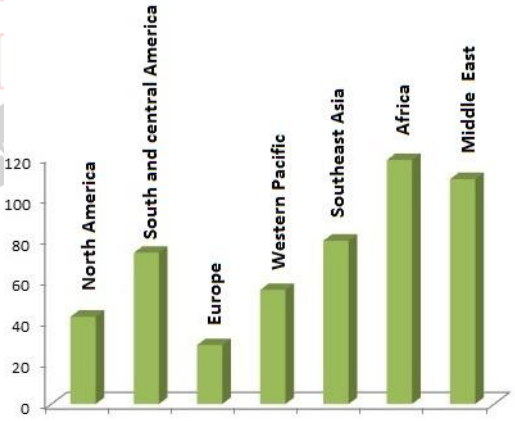


Figure 1 World Diabetic cases in 2035 approx

World health organization survey reports that about 1.5 million persons will die due to the diabetics and 2.2 million deaths will occur due to high blood glucose before 70 years of lifespan in 2030 as a approximate data. Based on this reports it is essential to conclude the diabetics death by analysing the data properly and to provide a proper medication is necessary in all over the world. Figure 1 gives an illustration of diabetics patients ratio in percentage by 2030. For this process the data from different sources is considered as heterogeneous in nature and it needs an application to assess the individual needs. A typical data

 DOI : 10.18231/2454-9150.2018.1169

pool contains more than 10000 items as a large scale data collection and response centre. Data clustering process is used to group information from heterogeneous groups and the important problem while developing clusters is due to the unavailability of information about the data set. The input parameters for developing cluster and its size based on the nearest neighbours makes the clustering into a challengeable task for a data set contains complex density, large size, noise contents. Figure 2 depicts the simple cluster analysis components starting from data set which is collected from medical examination details and then it is extracted based on the features. Extracted data contents is represented in similar categories and then the distance between the data points are identified for cluster analysis using suitable cluster mechanism. The results are represented using cluster validity index and the knowledge about the cluster and its contents is given to the user for their applications.

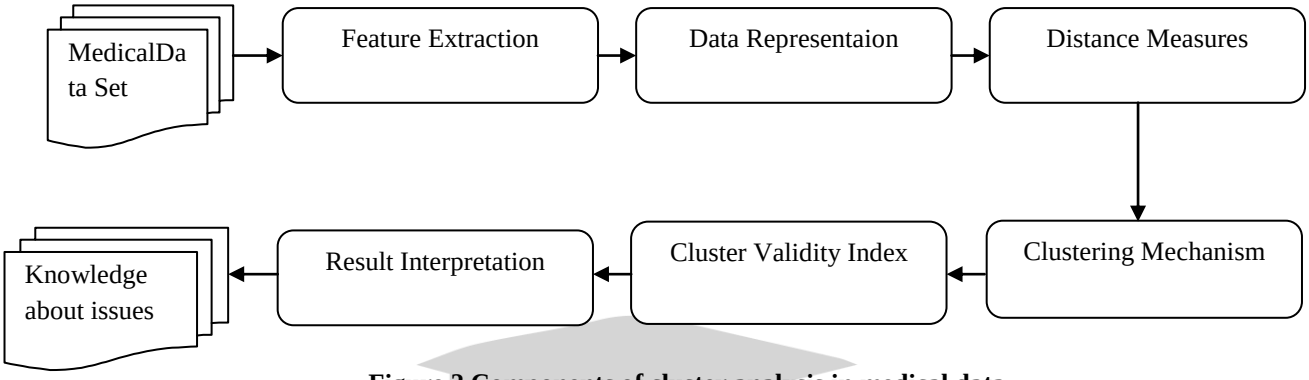


Figure 2 Components of cluster analysis in medical data

Based on the application, clustering techniques is classified and some clustering models considers the input parameters, amount of data, type of data to be clustered etc., some available clustering algorithms are portioning based clustering, Hierarchical based clustering, Model based clustering, density based clustering, grid based clustering, graph based clustering. Based on the parameters such as scalability, attributes to handle different data types, shape of the clusters, high dimensionality, noise data dealing and interpretability the clustering is defined for the dataset. Based on the objective function and compactness in minimizing the deviation while handling the intra-cluster data the proposed research work combines the spectral graph portioning with k-means algorithm and also it maximizes the connectedness in intra-clusters. Spectral graph portioning is a popular technique uses Eigen vectors of the Laplcian matrix which is constructed from the sample points obtained from samples of different groups. Based on parallel computations, load balancing different models of spectral based clustering is used. K-means clustering is a statistical based cluster analysis model to classify the given data set and the clusters compete with each other to own its data points. Is uses data points to assign into one cluster with equal contributions. The combination of this twp approaches is to map the high dimensional data using non linear function and then portioning the points using linear separators. Spectral graph partitioning is used to develop affinity matrix for clustered data. These two functions are related using the weights and its kernel functions. Using k-means the trace maximization problem can be solved and the suitable relaxation can be obtained using Eigen vectors in spectral graph portioning model.

II. RELATED WORKS

A survey has made in analyzing the issues present in existing clustering models, literature [1] describes about the data analysis model based on spectral graph partitioning for improving the accuracy. Based on the Eigen values and the analysis of upper bound values the issues in the approximation models are resolved in the model. Literature [2] provides a segmentation model using heterogeneous graph. It involves two graphs such as weighted graph and weighted dual mesh graph for clustering process. Using the laplacian values for the individual graph values the experimentations performed in the research work. Research work [3] provides a spectral based clustering process as hyper-graph spectral clustering and hyper-graph spectral clustering with local refinement. The nodes and its weights calculated for obtain the output and provide optimal results than existing models. It provides an efficient sketching algorithm for sub cluster, which is suitable for computer vision applications. Literature [4] reports about the detection mechanism used in diabetic patients data by implementing split and merge technique. The principle behind the algorithm uses local variation operator for fine detection. Combination of fine and coarse used in the proposed model provides better results in sensitivity and specificity. Literature [5] provides a detailed mathematical study about the ratio cut and normalized cut value in the spectral graph-partitioning model. It uses minimum cut partitions for the imbalanced clusters and develops parameters based on the cut analysis results. Research work [6] describes about the locally biased graph algorithm, which is applicable to small scale and large-scale data set. The issues in the traditional graph algorithm rectified using the proposed model and that is useful in downstream data applications. Literature [7] describes about the high order k means algorithm in the heterogeneous data model. Based on the tensor distance the research work represents heterogeneous datasets and achieve high clustering accuracy than the other clustering models. However, this high order data set lags in its performance for lower order data sets and time consumption is higher for large data sets. Research work [8] provides a detailed survey regarding spectral clustering in terms of its advantages and the issues.

 DOI : 10.18231/2454-9150.2018.1169

From this research work, the existing models such spectral model clustering and affinity propagation models considered for the comparison of proposed research work. Based on above survey work each research work lags in its performance in terms of clustering accuracy and overall time cost. Proposed spectral partitioning based k means clustering provides better accuracy and time cost than existing models and its mathematical model provided in the following section.

III. PROPOSED WORK

Based on the properties of Eigen vectors of the laplacian matrix the spectral graph partitioning function is defined by Donath, Hoffman, fiedler. The main objective of the spectral graph partitioning is to reduce the cut size between the two graphs which have same number of nodes in a heterogeneous dataset. Consider a laplacian matrix be $L = D - W$ and $|X| = |Y|$ and the cut size of the node is given as

$$C_s(X,Y) = cut(X,Y) \quad (1)$$

The indicator variables used in the model is $I_v = [1,-1]$ and it depends upon the nodesX,Y. The cut size including the indicator variable is given as

$$C_s(X,Y) = \sum_{v\in E} \frac{(I_u - I_v)^2}{4} W_{uv} \quad (2)$$

The value of I_u is a continues value from [-1,1] the Eigen system is solved into

$$(D - W)I = \lambda I \quad (3)$$

Since the trivial $I_1 = e$ is associated with $\lambda_1 = 0$, the second Eigen vector I_2 , the Fiedler vector, is the solution. The vector provides a good linear search for ratio cut portioning and it is given as

$$R_{cut} = \frac{Cut(X,Y)}{|X|} + \frac{Cut(X,Y)}{|Y|} \quad (4)$$

The use of General Eigen system and by using the normalized laplacian matrix the normalized cut value is given as

$$N_{cut} = \frac{Cut(X,Y)}{Deg(X)} + \frac{Cut(X,Y)}{Deg(Y)} \quad (5)$$

The $N_{cut}(X,Y)$ can be reduced into $J_N(Q)$. The same linear search based on the Q to obtain the minimum value of N_{cut} and the objective is given as

$$Min_Q N_{cut}(X,Y) \geq \rho_2 \quad (6)$$

The values ofR_{cut}, N_{cut} andM_{cut}, is considered based on the linear order vector and laplacian matrix for the appropriate search order. It is evident that the objective function is obtained based on the Eigen values of vectors using perturbation analysis over the normalized laplacian matrix. The value of M_{cut} is close to N_{cut} when the cut size is too small and the degree of sub graph split into its weights as

$$Deg(X) = \sum_{u\in X}\sum_{v\in Y} W_{u,v} = W(X) + Cut(X,Y) \quad (7)$$

And the N_{cut} can be modified as follows

$$N_{cut} = \frac{Cut(X,Y)}{W(X)+Cut(X,Y)} + \frac{Cut(X,Y)}{W(Y)+Cut(X,Y)} \quad (8)$$

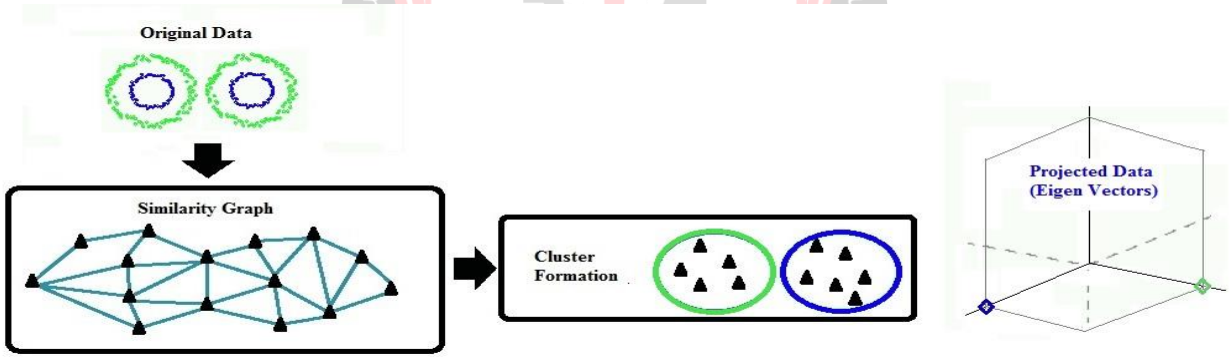


Figure 3 Illustration of Data clustering

When the $Cut(X,Y)$ overlaps between the two clusters the value becomes too small and then the M_{cut} is close to N_{cut}. If the $Cut(X,Y)$ is not small then there will be some substantial differences in the resultant sub graphs. If the overlap is large then any cluster can be identified easily.

Figure 3 gives an illustration of graph based cluster formation in a dataset. Other than spectral graph partitioning remaining models uses the singular value decomposition in the segmentation process. Some partitioning clustering is based on sum of squared error and based on the k means clustering the proposed model is defined. In the K means clustering the K data points is considered for initial cluster centres. It assigns each data point into a nearest centre and then forms the K clusters. Formed clusters are recomputed into cluster centres until the centre point is fixed. The vertex in the spectral graph portioning denotes the data point and the edge denotes the similarity between the points

 DOI : 10.18231/2454-9150.2018.1169

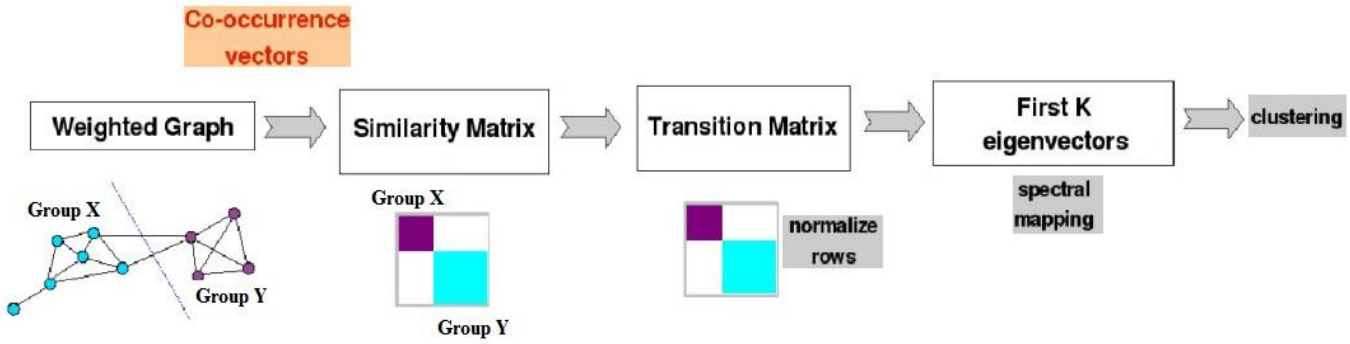


Figure 4 Proposed Clustering Model

The similarities produced by normalized cut have different sizes and shapes with different densities and it is not enough to form the cluster accurately. Transferring such data into the intrinsic structure based on the K means provides better performance. Figure 4 gives an illustration of proposed clustering model includes spectral graph partitioning and k-means algorithm. Applying K means to the Laplacian vectors helps to obtain the cluster even in the non-convex boundaries. The choice of K values in the cluster is most stable and usually it maximizes the Eigen distance and it is given as

$$\Delta_k = |\lambda_k - \lambda_{k-1}| \quad (9)$$

For minimizing the error function the model of K-means is given as

$$J = \frac{1}{n}\sum_{i=1}^{k}\sum_{I}|I - n_i|^2 \quad (10)$$

Where k is the given number of clusters, n_i is the prototype of cluster

The proposed algorithm is described as follows

Input: A set of data points and K number of Clusters
Output: Clusters $V_1, V_2 \ldots . V_K$
Construct Similarity Graph of data points and the corresponding affinity matrix weight is taken as W
Compute L=D-W
Compute the First k Eigen vectors
Construct the matrix of $n \times k$
Consider each row of matrix as data points and cluster the points using K means Algorithm

---

--Step 1: K means Algorithm
Spilt the data set into clusters
Select the K highest degree nodes from initial values
Apply K means and produce partitions
For each partition, find the main node
For each sub node repeat the process and combine until it belongs to main node

---

Merge the nodes into final clusters
Generate the Neighbor pairs
For each pair calculate the merge criteria value
Repeat until it gets maximum pair value as K cluster have been obtained
Add the pairs to the main cluster

Stop the process until it Converges
Else
Repeat step 1

IV. RESULT AND DISCUSSION

The proposed model is experimented AIM '94 data set and the process is performed without any prior knowledge about the dataset to achieve better clustering accuracy. The value of k considered into 5 for the dataset and the same value is used for spectral clustering and affinity propagation. Experimentation performed in an Intel i3 processor at 1.87GHz with 4GB of RAM. The data set contains 400 data points distributed into three separate clusters. The performance of clusters can simply obtained using K means but the proposed spatial graph partitioning based clustering model detects the clusters accurately and figure 5 shows the results of clustering using proposed model. The input in the spectral clustering is based on the user definition and sigma values.

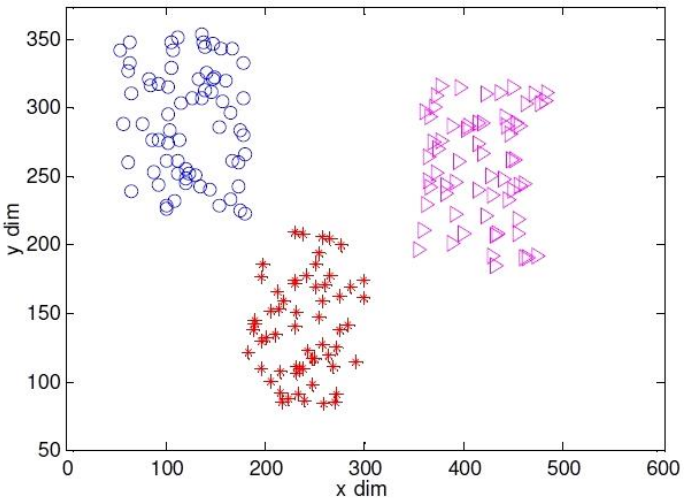


Figure 5 Clustering Results our proposed model (using k=5)

The value of k is taken into different levels to compare the performance level of sub clusters. Increasing the value of k reduces processing time as the shrinking process reduces and the complexity of the group decreases. Figure 6(a) and 6(b) depicts the clustering results of spectral model and affinity propagation model where it does not use any input parameters. In this it is visible that spectral model has better clustering accuracy than other by seeing its clustering process.

 DOI : 10.18231/2454-9150.2018.1169

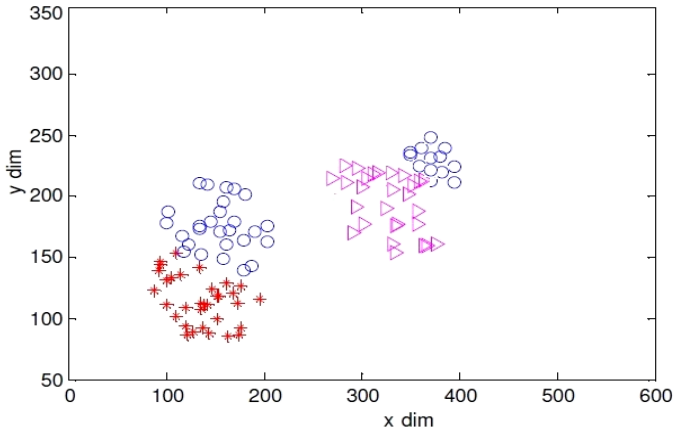
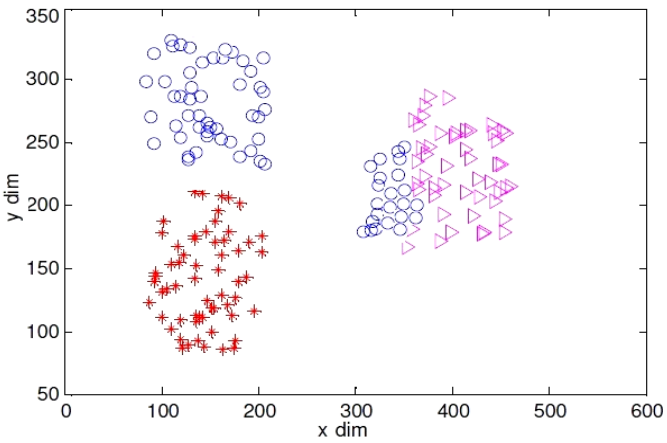


Figure 6 (a) Affinity Propagation Clustering



6 (b) Spectral clustering method

In figure 7 the overall time cost for the proposed model and the existing model is compared by varying the k values. As discussed earlier the time cost will be reduced as the value of k increases. It is observed that proposed model has less time cost for entire process. This is achieved by using standard k values starting from 0 to 80. Compared to affinity propagation spectral model provides less time cost but it lags in its performance as the range reaches 60-70. This is overcome using proposed model and it provides an average time cost for all the ranges of k values. Figure 8 provides the clustering accuracy plot for the proposed model for each value. It is observed that proposed model has higher steady state accuracy for majority of the k values. For the values of k=60, 70, 80 the clustering accuracy is 91.26%, 93.46%, 95.42%. Also it is observed that existing model has lesser classification accuracy than the proposed model.

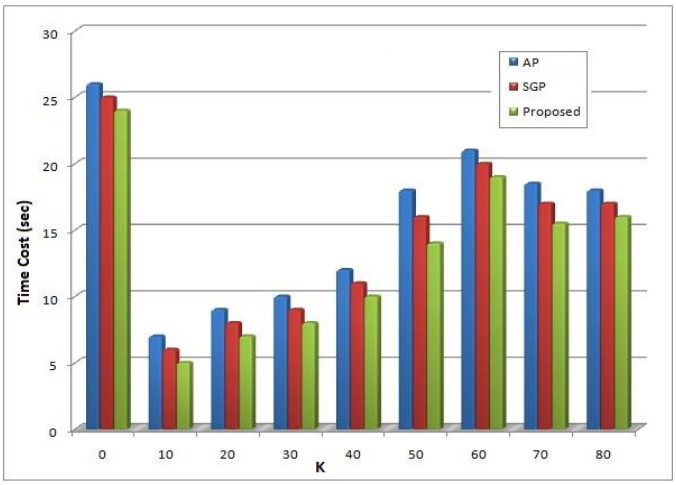


Figure 7 Overall time cost

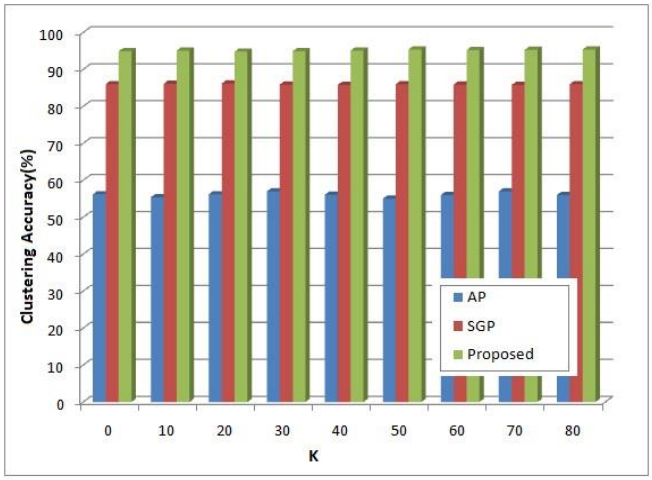


Figure 8 Clustering Accuracy

The effectiveness and efficiency of the proposed model is validated based on the overall time cost and clustering accuracy for the entire model using the k value as 5 and table 1 gives the comparison of proposed model with other models and the time to complete the clustering process is calculated in seconds.

Table 1 Performance comparison of Proposed model

S.No	Algorithm	Time cost(sec)	Clustering Accuracy (%)
1	Proposed SGP K means (k=5)	0.26	95.42
2	Spectral graph partitioning	0.42	86.94
3	Affinity Propagation	3.54	56.21

As shown in Table 1, our proposed algorithm has a clustering accuracy as high as 95.42% for the datasets that have been used. For the same data set, we have an accuracy of 86.94% for the spectral clustering method and 56.21 for affinity propagation model. The exact number of clusters present in dataset that included in all the models and the proposed model outperforms it in the execution time. The proposed model acquires less time cost and increased clustering accuracy for the entire data set.

V. CONCLUSION

The proposed spectral graph portioning using k means divides the clusters and provides better results for the entire data set. The vertices is obtained for the sub clusters and the simulation results shows that proposed model has better

 DOI : 10.18231/2454-9150.2018.1169

clustering accuracy and time cost than the other models such as spectral model and affinity propagation model. The proposed model has better accuracy of 95.42% which is 30% greater than the existing model in analysing the diabetic patient dataset.

REFERENCE

[1] James P. Fairbanks, David A. Bader, Geoffrey D. Sanders, "Spectral partitioning with blends of eigenvectors" Journal of Complex Networks, Vol.5 , No.4, Pp.551 – 580, 2017

[2] Panagiotis Theologou, Ioannis Pratikakis, Theoharis Theoharis, "Unsupervised Spectral Mesh Segmentation Driven by Heterogeneous Graphs" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.39, No.2, Pp.397 – 410, 2017

[3] Kwangjun Ahn , Kangwook Lee, Changho Suh, "Hypergraph Spectral Clustering in the Weighted Stochastic Block Model" IEEE Journal of Selected Topics in Signal Processing, Vol.12, No.5,Pp. 959 – 974, 2018

[4] Cem Aksoylar, Jing Qian, Venkatesh Saligrama, "Clustering and Community Detection With Imbalanced Clusters" IEEE Transactions on Signal and Information Processing over Networks, Vol. 3 , No.1,Pp. 61 – 76, 2017

[5] Hussain F. Jaafar, Asoke K. Nandi and Waleed Al-Nuaimy, "Automated Detection of Exudates in retinal Images Using a Split and Merge Algorithm",European Signal Processing Conference, EUSIPCO, 2010

[6] Kimon Fountoulakis, David F. Gleich, Michael W. Mahoney, "An Optimization Approach to Locally-Biased Graph Algorithms" Proceedings of the IEEE, Vol.105 , No.2, Pp. 256 – 272, 2017

[7] Fanyu Bu, "A High-Order Clustering Algorithm Based on Dropout Deep Learning for Heterogeneous Data in Cyber-Physical-Social Systems" IEEE Access, Vol.6, Pp.11687 – 11693, 2018

[8] Cuimei Guo ; Sheng Zheng ; Yaocheng Xie ; Wei Hao, "A survey on spectral clustering" World Automation Congress 2012, IEEE, 2012

 DOI : 10.18231/2454-9150.2018.1169