Research Journal of Applied Sciences, Engineering and Technology 8(20): 2097-2102, 2014
DOI:10.19026/rjaset.8.1203
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: June 08, 2014 Accepted: July 13, 2014 Published: November 25, 2014

Research Article

Modified Structural and Attribute Clustering Algorithm for Improving Cluster Quality in Data Mining: A Quality Oriented Approach

[1]G. Abel Thangaraja and [2]Saravanan Venkataraman Tirumalai
[1]Department of Computer Science, Kaypeeyes College of Arts and Science, Kotagiri, the Nilgiris, India
[2]College of Computer and Information Sciences, Majmaah University, Majmaah,
Kingdom of Saudi Arabia

Abstract: The need of Data mining is because of the explosive growth of data from terabytes to petabytes. Data mining preprocess aims to produce the quality mining result in descriptive and predictive analysis. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. A straightforward way to combine structural and attribute similarities is to use a weighted distance function. Clustering results are arrived based on attribute similarities. The clusters balance the attribute and structural similarities. The existing Structural and Attribute cluster algorithm is analyzed and a new algorithm is proposed. Both the algorithms are compared and results are analyzed. It is found that the modified algorithm gives better quality clusters.

Keywords: Attribute similarity, cluster quality, data mining, structural

INTRODUCTION

Data mining: The need of Data mining is because of the explosive growth of data from terabytes to petabytes. As databases grows larger, decision-making from the data is too complicate so we need data mining to derive knowledge from the stored data. Data mining is also called as Knowledge Discovery (mining) in Databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc. Data mining preprocess aims to produce the quality mining result in descriptive and predictive analysis (Jeyabalaraja and Edwin Prabakaran, 2012). Data mining techniques are used in different application to analysis and predict the data for decision support system. Data mining refers to extracting or mining knowledge from large amounts of data.

Cluster quality: A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns (Han *et al*., 2011).

Structural similarity: Clusters who are having only structures which give the outcomes based on vertex connectivity are called as Structural similarity. Similarity is expressed in terms of a distance function. A straightforward way to combine structural and attribute similarities is to use a weighted distance function. Distances are normally used to measure the similarity or dissimilarity between two data objects.

Attribute similarity: Clustering results are arrived based on attribute similarities. The clusters balance the attribute and structural similarities. Vertex distances and similarities have been measured by random walk principle. A unified framework based on Neighbourhood random walk is to integrate structural and attribute similarities.

The clusters balance the attribute and structural similarities. Vertex distances and similarities have been measured by random walk principle. The purpose of this problem is to partition the attributed graph into k clusters with intracluster attributes. This partitioning is complicated because attributed and structural similarities are independent. In this study, consider a dataset of scores obtained by university students in two subjects. Each pair represents the vertex of the graph. The techniques adopted in this study are listed below:

- Propose a unified Neighbourhood random walk distance measure to combine attribute and structural similarities.
- Theoretical methods are given to boosten the presentations of attribute similarity to the unified

Corresponding Author: G. Abel Thangaraja, Department of Computer Science, Kaypeeyes College of Arts and Science, Kotagiri, the Nilgiris, India

Neighbourhood random walk distances for studying the closeness of the vertices.
- Apply a weight self-adjustment method to analyze the degree of contributions of attributes in random walk distances.
- Perform suitable experiments using designed clustering algorithm.

LITERATURE REVIEW

Graph clustering techniques have been analyzed in many directions and primarily concentrated on topological structures. Strehl and Ghosh (2002), have studied Ensemble analysis which improves classification accuracy and the general quality of cluster solution. They have also discussed the availability of multiple segmentation solutions within an ensemble and the method is Meta clustering algorithm and is based on the notion of "clustering on clusters". Pons and Latapy (2006), have proposed short random walks of length 'l' to measure the similarity between two vertices in a graph for community detection. Tsai and Chui (2008), have developed a feature weight self-adjustment mechanism for k-means clustering on relational datasets. Here, an optimization model is designed to find feature weights in which the partitions within clusters are minimized and that between clusters are maximized.

Orme and Johnson (2008) have discussed ensemble analysis for improving k-means cluster analysis and the methods have been described with the help of numerical illustrations. Zhou *et al*. (2009), have proposed graph clustering algorithm based on both structural and attribute similarities and estimated the effectiveness of SA cluster as compared with other three clusters, through experimental analysis. Rai and Singh (2005) have summarized and described the types of clusters and different clustering methods.

Zanghi *et al*. (2009), have adopted generative process and proposed a probabilistic model to cluster attributed graphs. Tajunisha and Saravanan (2011), have proposed a method to find initial centroid for k-means and they have used similarity measure to find the informative genes. The goal of their clustering approach is to perform better cluster discovery on sample with informative gene. Cheng *et al*. (2011), have studied graph clustering using unified random walk distance measures. A comparative analysis of clusters and their efficiencies have been carried out.

Ghasemi *et al*. (2013) have presented a new procedure to find functional modules in PPI networks. The main idea is to model a biological concept and to use this concept for finding good functional modules in PPI networks. In order to evaluate the quality of the obtained clusters, the results of proposed algorithm is compared with those of some other widely used clustering algorithms.

Wang and Davidson (2010), have proposed a research work as a natural extension to unconstrained spectral clustering and are interpreted as finding the normalized min-cut of a labeled graph. The effectiveness of this approach by empirical results on real-world data sets, with applications to constrained image segmentation and clustering benchmark data sets with both binary and degree-of-belief have been validated. Jayabrabu *et al*. (2012) have analyzed the formulated clusters quality based on quality parameters by using Data mining agents. Clustering algorithms will produce clusters based on given input data. But, it is noted that all clusters are not good clusters.

METHODOLOGY

Distance measures: The distance measure is defined as the distance between two objects O_1 and O_2 universe of objects denoted as d (O_1, O_2) which non-negative real number is always. Distance measure are use to obtain the similarity or dissimilarity between any pair of objects. In general, distance measures are used for Numeric attributes (Minkowski metric) (Han *et al*., 2011), Binary attributes, Nominal attributes, Ordinal attributes and Mixed type attributes.

Unified neighbourhood random walk distance: Let P_A be the transition probability matrix of augmented graph G_a and it is formed by using the transition probabilities from vertices γ_i to γ_j through attribute and structure edges.

Given the length of the random walk as 'l' with the probability of restart $c \in (0, 1)$. The unified Neighbourhood random walk distance $d(\gamma_i, \gamma_j)$ from γ_i to γ_j in G_A is defined as:

$$d(\gamma_i, \gamma_j) = \sum_{\substack{\tau: \gamma_i \to \gamma_j \\ \delta \le l}} P_A(\tau)\, c\, (1-c)^{\delta} \quad (1)$$

where, τ is the path from γ_i to γ_j whose length is denoted as δ with transition probability $P_A(\tau)$. The Eq. (1) can be written in matrix form as:

$$R_A^l = \sum_{r=0}^{l} c\, (1-c)^r\, {P_A}^r \quad (2)$$

Here, R_A is the Neighborhood random walk distance matrix.

Clustering process: Clustering process has the duty of separating the data into different clusters with same or different characters. The selection of good initial centroid is more powerful than that of randomly selected initial centroids.

In order to select the centroids, define the density function of vertex.

The density function of a vertex γ_i is the sum of the influence functions of γ_i on all vertices in V.

The influence function is stated as:

$$f_B^{\gamma_j}(\gamma_i) = 1 - e^{-\frac{1}{2\sigma^2}\{d(\gamma_i,\gamma_j)\}^2} \quad (3)$$

Hence, the density function is written as:

$$f_B^D(\gamma_i) = \sum_{\gamma_j \epsilon V}\left[1 - e^{-\frac{1}{2\sigma^2}\{d(\gamma_i,\gamma_j)\}^2}\right] \quad (4)$$

It is noted that the influence of γ_i on γ_j is proportional to the random walk distance from γ_i to γ_j. We know that larger random walk distance gives more influence. If γ_i has a large density value, then γ_i connects to many vertices.

By using the density functions given in "(4)", the vertices are arranged in descending order of their densities and select the top k vertices whose initial centroids are stated as $\{c_1^0, c_2^0, \dots, c_k^0\}$. After a large number of iterations are performed, the k centroids in the t^{th} iteration are $\{c_1^t, c_2^t, \dots, c_k^t\}$.

Consider ω_0 is the initial weight of structure edge and $\omega_1, \omega_2, \dots, \omega_m$ are the initial weights of attribute edge which are relative to ω_0. Assuming $\omega_0^0 = 1.0$ and $\omega_1^0 = \omega_2^0 = \cdots = \omega_m^0 = 1.5$.

Let $W^t = \{\omega_1^t, \omega_2^t, \dots, \omega_m^t\}$ be the attribute weights in the t^{th} iteration. An increment $\Delta\omega^t$ is weight update of attribute a_i between the t^{th} and $(t+1)^{th}$ iterations. The weight of a_i in the $(t+1)^{th}$ iteration is defined as the average of weight in the t^{th} iteration and its increment. That is:

$$\omega_i^{t+1} = \frac{1}{2}(\omega_i^t + \Delta\omega_i^t) \quad (5)$$

The expression for the increment of the weight in the t^{th} iteration is used in "(5)" which gives:

$$\omega_i^{t+1} = \frac{1}{2}\left[\omega_i^t + \frac{m\sum_{j=1}^{k}\sum_{\gamma\epsilon V_j} vote_i(c_j,\gamma)}{\sum_{p=1}^{m}\sum_{j=1}^{k}\sum_{\gamma\epsilon V_j} vote_p(c_j,\gamma)}\right] \quad (6)$$

where,

$$vote_i(\gamma_p,\gamma_q) = \begin{cases} 1, & \text{if vertices share the same value on } a_i \\ 0, & \text{otherwise} \end{cases}$$

EXISTING STRUCTURAL AND ATTRIBUTE CLUSTER APPROACH

The following algorithm is used to evaluate cluster centroids and adjusted weights for different iterations. This clustering result balances the structural and attributes similarities.

Algorithm: Attributed Graph Clustering Structural and attribute Cluster.

Input: An attributed graph G, a length limit l of random walk paths, a restart probability c, a parameter σ of influence function, cluster number k.

Output: k clusters $V_1, \dots, V_k$:

- Initialize $\omega_1 = \omega_2 = \cdots = \omega_m = 1.5$ fix $\omega_0 = 1.0$
- Calculate the unified random walk distance matrix $-R_A^l$
- Select k initial centroids with highest density values
- Repeat until the objective function converges
- Assign each vertex γ_i to a cluster C^* where the centroid $c^* = \arg\max_{c_j} d(\gamma_i, c_j)$
- Update cluster centroids with the most centrally located point in each cluster using random walk distance vector
- Update weights $\omega_1, \dots, \omega_m$ in large number of iterations
- Re-calculate R_A^l
- Return k clusters $V_1, \dots, V_k$

Structural and Attribute Cluster (SAC) algorithm considers both structural and attribute similarities. The input parameters for the SAC algorithm are as follows: an attributed graph G, a length limit l of random walk paths, a restart probability c, a parameter σ of influence function and cluster number k. In step 1, consider the Intial weights as equal i.e., $\omega_1 = \omega_2 = \cdots = \omega_m = 1.5$ fix $\omega_0 = 1.0$. In step 2, measure the unified random walk distances between the vertices and construct unified random walk distance matrix R_A^l. In step 3, the density functions of the vertices are calculated and choose the highest density value which is the key to select the initial centroid. By doing this process, k initial centroids are obtained. In step 4, perform a large number of iterations, the k centroids in the t^{th} iteration are obtained as $\{c_1^t, c_2^t, \dots, c_k^t\}$. In step 5, assign each vertex to the closest centroid with largest random walk distance from γ_i. In step 6, use random walk distance vectors and its average, the cluster centroids are updated with the most centrally situated vertex in each cluster. In step 7, the weights of the ith attribute are obtained in large number of iterations. In step 8, recalculate the unified random walk distances between the vertices and construct unified random walk distance matrix R_A^l. In the last step the required k clusters are evaluated as $V_1, \dots, V_k$.

Modified SA cluster: Modified Structural and Attribute Clustering (MSAC) algorithm is an advancement of SA clustering technique. SAC is used only for graph based clustering. The proposed MSAC can be used for any type of data. The input for MSAC, i.e., C_1, C_2.....C_k is the set of clusters generated from DBSCAN algorithm. In step 1, the values of the structural and attribute similarity are calculated for the clusters $C_1, \dots, C_k$. In step 2, centroids are chosen in random for each and every cluster with low similarity and attribute value. In step 3, the members with high similarity and attribute value are reallocated to another cluster so as to maintain the low value. This process is repeated for all the cluster members $C_1, \dots C_K$. In step 6, the cluster members in each cluster are checked for

the distance between the centroid and the member with less similarity and attribute value in each cluster. The quality clusters $Q_1, \ldots, Q_k$ are obtained finally.

Algorithm Modified SA Cluster.
Input: Set of clusters $C_1, \ldots . C_k$ from DBSCAN.
Output: k clusters $Q_1, \ldots, Q_k$:

1. Calculate the structural and attribute similarity for $C_1, \ldots C_k$
2. Select k initial centroids with low similarity and attribute value
3. Converge the clusters with high similarity and attribute value to low value
4. Repeat for all the cluster members $C_1, \ldots . C_K$
5. If needed, reposition the cluster members
6. Update cluster centroids with less similarity and attribute value in each cluster
7. Return k clusters $Q_1, \ldots, Q_k$

Database used: A student database was used in the existing SAC algorithm and proposed Modified SAC algorithm. The student details table consists of 2 attribute and 5547 rows. The 2 attributes are scores of Subject I and subject II.

RESULTS AND DISCUSSION

As stated in Section above, consider 50 sets of scores of students in two subjects which are dependent with each other. Each pair of scores is considered as vertex of a graph. After fixing the vertices and their edges in the two dimensional graph, by using Structural and Attribute Clustering (SAC) Algorithm, the Adjacency matrix (Adjacency matrix), Transition probability matrix (Transition probability matrix) and Neighborhood Random walk distance matrix (Neighborhood random walk distance matrix) are found. Subsequently, on applying distance matrix in the influence functions which give density functions (Densities). These results are given in the Appendix. Finally, the vertices are grouped into several clusters. It is observed that the vertices are grouped into 5 clusters (Fig. 1) which are produced by SAC Algorithm.

In other way, the number of clusters is easily found. When Modified Structural and Attribute Clustering (MSAC) Algorithm is adopted to classify the vertices into clusters, DBSCAN plays a main role to give the number of clusters of the vertices. After obtaining the number of clusters, the adjacency matrix and the subsequent measures until the density functions are obtained. These results exhibit 6 clusters (Fig. 2). On comparing both Algorithms, MSAC is better than SAC, since MSAC identified more number of clusters and quality clusters than that of SAC.

CONCLUSION

The results for the modified structural and attribute clustering algorithm show that the cluster is of good quality when compared with the existing SA cluster. In the SAC algorithm, the k value has been given by the user, but in MSAC algorithm, the k value has been calculated by DBSCAN. It is concluded that MSAC is more effective than the existing Algorithm.

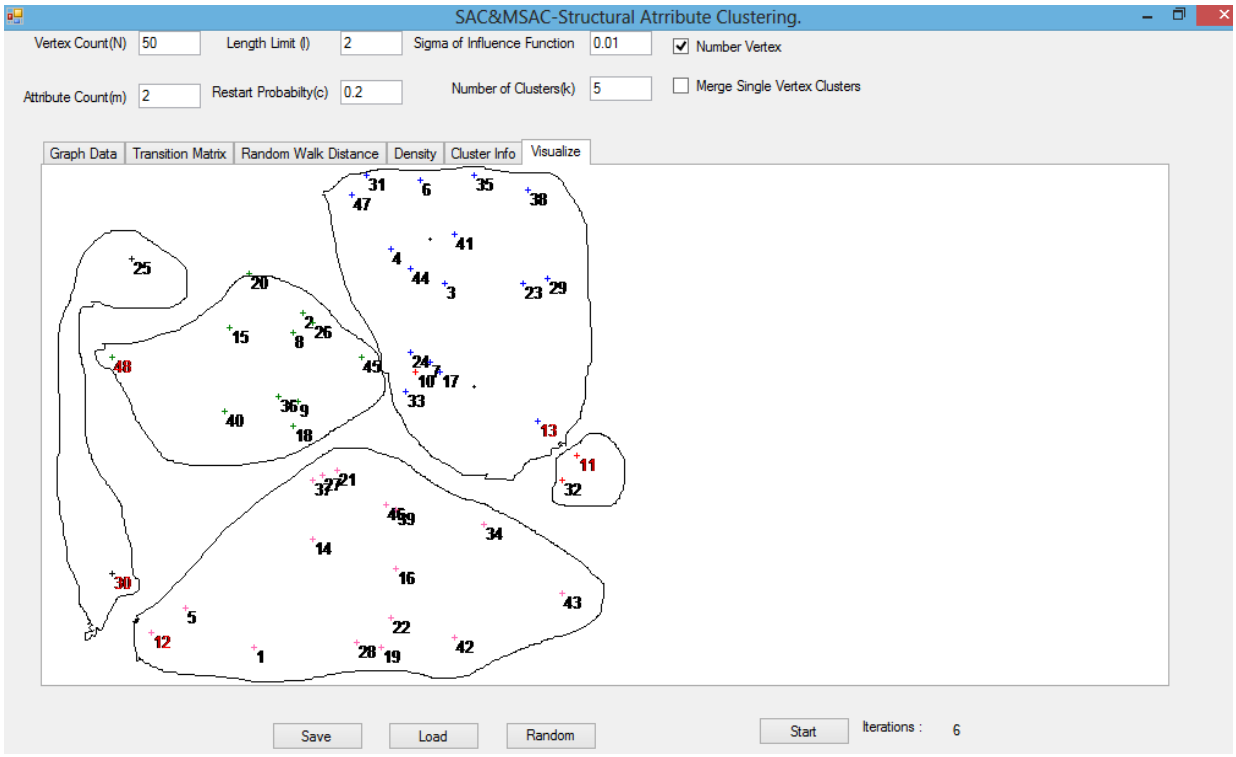

Fig. 1: Clustering by SAC algorithm

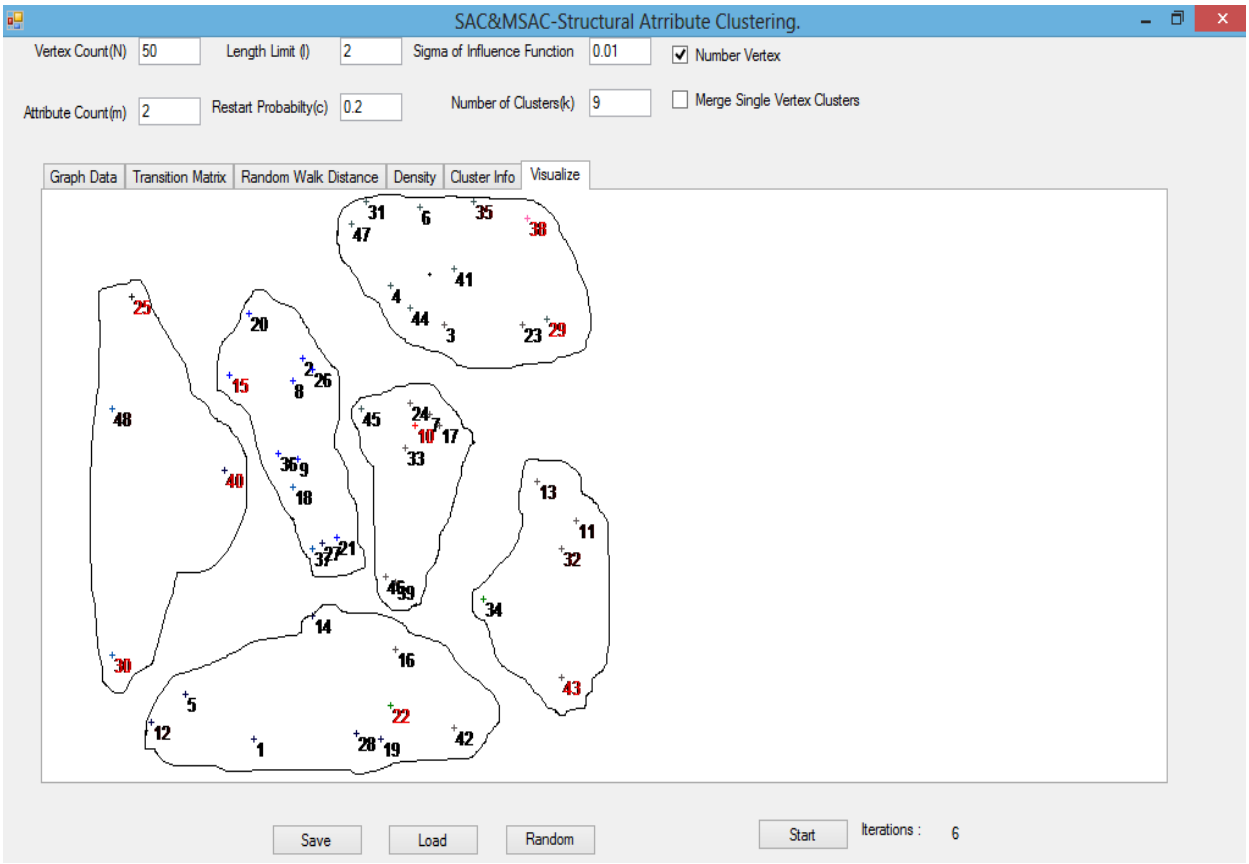

Fig. 2: Clustering by MSAC algorithm

APPENDIX

Adjacency matrix: A (50×50)

$$\begin{bmatrix} 0 & 0 & 0 & & 0 & 0 & 0 \\ 0 & 0 & 1 & \cdots\ \cdots & 1 & 1 & 1 \\ 0 & 1 & 0 & & 0 & 1 & 1 \\ & \vdots & & \ddots & & \vdots & \\ 0 & 1 & 0 & & 0 & 0 & 0 \\ 0 & 1 & 1 & \cdots\ \cdots & 0 & 0 & 1 \\ 0 & 1 & 0 & & 0 & 1 & 0 \end{bmatrix}$$

Transition probability matrix: $P_A(\tau)$ (50×50)

$$\begin{bmatrix} 0 & 0 & 0 & & 0 & 0 & 0 \\ 0 & 0 & 0.019231 & \cdots\ \cdots & 0.019231 & 0.019231 & 0.019231 \\ 0 & 0.019231 & 0 & & 0 & 0.019231 & 0.019231 \\ & \vdots & & \ddots & & \vdots & \\ 0 & 0.019231 & 0 & & 0 & 0 & 0 \\ 0 & 0.019608 & 0.019608 & \cdots\ \cdots & 0 & 0 & 0.019231 \\ 0 & 0.019608 & 0.019608 & & 0 & 0.019231 & 0 \end{bmatrix}$$

Neighborhood random walk distance matrix: R_A (50×50)

$$\begin{bmatrix} 0.20071 & 0.200331 & 0.200237 & & 0.200142 & 0.200379 & 0.200047 \\ 0.200331 & 0.201564 & 0.204357 & \cdots\ \cdots & 0.20355 & 0.20426 & 0.204118 \\ 0.200237 & 0.204357 & 0.201564 & & 0.200379 & 0.204402 & 0.20426 \\ & \vdots & & \ddots & & \vdots & \\ 0.200142 & 0.20355 & 0.200379 & & 0.200568 & 0.200331 & 0.200379 \\ 0.200386 & 0.204344 & 0.204489 & \cdots\ \cdots & 0.200338 & 0.201641 & 0.20111 \\ 0.200048 & 0.204199 & 0.204344 & & 0.200386 & 0.20111 & 0.201255 \end{bmatrix}$$

Densities (50): 14.79778, 30.98239, 30.99801, 26.99986, 16.8756, ... , 30.99269, 23.935, 11.50352, 0, 0.

REFERENCES

Cheng, H., Y. Zhou and J.X. Yu, 2011. Clustering large attributed graphs: A balance between structural and attribute similarities. ACM T. Knowl. Discov. Data, 5(2), Article12.

Ghasemi, M., M. Rahgozar, G. Bidkhori and A. Masoudi-Nejad, 2013. *C-element*: A new clustering algorithm to find high quality functional modules in PPI networks. PLoS one 8(9), DOI: 10.1371/journal.pone.0072366.

Han, J., M. Kamber and J. Pei, 2011. Data Mining: Concepts and Techniques. 3rd Edn., Elsevier Science, Burlington.

Jayabrabu, R., V. Saravanan and K. Vivekanandan, 2012. A framework: Cluster detection and multidimensional visualization of automated data mining using intelligent agents. Int. J. Artif. Intell. Appl., 3(1): 15.

Jeyabalaraja, V. and T. Edwin Prabakaran, 2012. Study on software process metrics using data mining tool-A rough set theory approach. Int. J. Comput. Appl., 47(18): 0975-888.

Orme, B. and R. Johnson, 2008. Improving K-means cluster analysis: ensemble analysis instead of highest reproducibility replicates. Proceeding of the Saw Tooth Software Conference. Sequim WA.

Pons, P. and M. Latapy, 2006. Computing communities in large networks using random walks. J. Graph Algorithm. Appl., 10(2): 191-218.

Rai, P. and S. Singh, 2005. A survey of clustering techniques. Int. J. Comput. Appl., 7(12): 1-5.

Strehl, A. and J. Ghosh, 2002. Clustering ensembles-A knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res., 3(2002): 583-617.

Tajunisha, N. and V. Saravanan, 2011. A new approach to improve the clustering accuracy using informative genes for unsupervised microarray data sets. Int. J. Adv. Sci. Technol., 27: 85-94.

Tsai, C.Y. and C.C. Chui, 2008. Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm. Comput. Stat. Data An., 52: 4658-4672.

Wang, X. and I. Davidson, 2010. Flexible constrained spectral clustering. Proceeding of 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10), pp: 563-572.

Zanghi, H., S. Volant and C. Ambroise, 2009. Clustering based on random graph model embedding vertex features. Pattern Recogn. Lett., 31(9): 830-836.

Zhou, Y., H. Cheng and J.X. Yu, 2009. Graph clustering based on structural/attribute similarities. Proceedings of the VLDB Endowment, 2(1): 718-729.