International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 12, December-2014

Proposed Enhancement in the Performance of the Spectral Clustering

Gurpinder Kaur
Department of computer science and engineering
Lovely Professional University, Phagwara
Punjab, India

Mr. Abhishek Tyagi
Department of Computer Science and engineering
Lovely Professional University, Phagwara
Punjab, India

***Abstract*- Spectral clustering has become one of the most hotspots in clustering over the past few years. It has been proved to be effective by the many researchers and its performance outperforms many other clustering techniques. Much work has been done to enhance the performance of the spectral clustering but still many of them are quite computationally expensive. In this work, focus will be on reducing the computation time and also on the accuracy in order to improve the quality of clusters. The base paper used the Gaussian kernel functions to construct the similarity matrix and in this work some other light weight functions will be used instead of the Gaussian kernel functions and it is expected to be effective in achieving the expected goals.**

Keywords*- *Spectral clustering, computation time, quality of clusters.

I. INTRODUCTION

Data mining (also known as data or knowledge discovery process) is a process of evaluating useful data from a large amount of data by analyzing the data from different perspectives [5].

It is a process of extracting useful, meaningful and relevant data from a huge amount of data. Clustering is one of the most widely used techniques for analysing the data which attempts to keep similar kind of data together and dissimilar data apart from each other. Data clustering is a method in which the whole data under consideration is divided into clusters and this division process depends upon the characteristics of the data. The data with similar characteristics is kept in one cluster and those with different are kept in different clusters. Hence the main motive of the clustering is to maximize the intra-cluster similarity and minimize the inter-cluster similarity. The major application areas of cluster analysis include market research, pattern recognition, data analysis, image processing and outlier detection applications such as detection of credit card fraud.

There exists many different types of clustering methods including hierarchical, partitioning, density-based, model-based and grid-based methods.

Hierarchical methods (also known as connectivity-based methods) use the approach of recursively partitioning the instances in either a top-down fashion or bottom-up fashion in order to form the clusters. Depending on this top-down and bottom-up fashion, this method can be further divided into two types- agglomerative hierarchical and divisive hierarchical.

Agglomerative is a bottom-up approach in which each object initially represents a cluster of its own and then they are successively merged until the desired cluster structure is obtained. Divisive-hierarchical clustering uses top-down approach where initially all the instances are in one single cluster which is then divided into sub-clusters and then these sub-clusters are successively divided into sub-clusters until the desired cluster structures are obtained [2]. The major disadvantage of hierarchical methods is that any action (split or merging) once performed can never be undone [8].

Partitioning methods (also known as centroid-based methods) relocate instances by moving them from one cluster to another. Such methods typically demand the number of required clusters to be determined in advance.[2] The most common example of this clustering method is k-means clustering which requires the number of clusters to be determined in advance and the same number of data points are selected to act as centroids.

In Density-based clustering, the density of the neighbourhood instances is checked to form the clusters. The areas of higher density are defined as clusters. The most popular density based clustering method is DBSCAN. It is based on connecting points with certain threshold distance [6].

Model-based clustering is also known as distribution based clustering. This method assumes that the data were generated by model and tries to recover the original model from the data. The model that is recovered from the data then defines clusters and assignment of documents to clusters [7].

In the grid-based method, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure. The major advantage of this method is its fast processing time [8].

Spectral clustering is a form of density-based based clustering method in which all the data points of the data set are represented in the form of nodes and the relationships between them are represented with the help of edges which carry some weight.

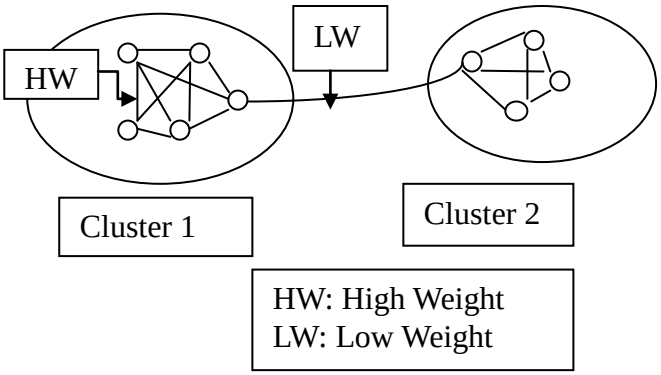

Figure 1: Basic Overview of Spectral Clustering

High weight indicates high similarity between the two nodes and the low weight is an indication of the dissimilarity between the two nodes. These weights between the nodes are recorded in the form of a matrix and the matrix so produced is known as distance or weight matrix. This distance matrix is then converted into similarity matrix using some conversion method. After this, laplacian matrix is constructed which is usually the difference between the distance matrix and the similarity matrix, the next step is to calculate the eigenvalues and construct a matrix that consists of these eigenvlaues as the columns. Then cluster the data points using some clustering algorithm.

II. LITERATURE REVIEW

Hao Huang *et al.* (2014) described that it is an open challenge to mine the arbitrary shaped clusters in large data sets. Various approaches to this problem have been proposed but they all have very high time complexity. In order to save the computational cost, some algorithms have made attempts to shrink/collapse the size of the data set to a smaller amount of representative data examples. However, this kind of user-defined shrinking ratios may significantly affect the performance of the clustering. In this paper, they present CLASP, by adopting three phase strategy that has been proved to be an effective and efficient algorithm for mining the clusters having an arbitrary shape. The first phase of their approach attempts to shrink the size of a data set automatically while effectively retaining the information about the shape of clusters in the data set with representative data examples. Then in the second phase, it attempts to adjust the positions of these representative data examples to increase their intrinsic relationship and make the structures of the clusters more clear and distinct for clustering. Finally in the third and last phase, it performs agglomerative clustering to identify the structure of the cluster with the help of a mutual k-nearest neighbors-based similarity metric called Pk for completing the mining of arbitrary shaped clusters [11].

Sumuya Borjigin and Chonghui Guo (2013) proposed non-unique cluster number determination methods based on stability. They used Gaussian kernel parameters (global scale and local scale) to convert the distance matrix into the similarity matrix and then used the multi-way normalized cut algorithm to cluster the data points. In their work they also focused on determining whether the chosen cluster numbers are stable and reasonable which is helpful in improving the performance of the clustering procedure. Also the coherence is measured based on the gap to measure clustering quality.

Xianchao Zhang and Quanzeng You (2010) proposed a random walk based approach to process the Gaussian kernel similarity matrix. In order to make the similarity matrix close to the ideal matrix, the pairwise similaity between two data points they consider was not only related to two data points, but also related to their neighbours. To deal with the noisy items, initially the noisy items were ignored and only the other items were taken into account to perform the clustering. After clustering the non-noisy items, the correct cluster for each noisy item is determined.

Cuimei Guo et al. (2010) discussed the basic framework of spectral clustering. They introduced the basic theories relative to spectral clustering and some algorithms also. Some of the partition criterions they included in their paper include minimum cut, normalized cut and multiway normalized cut. Besides this, they also focussed on the problems related to the spectral clustering which includes (a) constructing the efficient similarity matrix and the graph laplacian, (b) to decide the parameters of spectral clustering such as number of clusters and sigma and (c) extending spectral clustering to large data sets.

Xu-Degang et al. (2009) focused on building the affinity matrix, which is the most important part of spectral clustering and affects the process and quality of spectral clustering to a great extent. They proposed four different methods to build the affinity matrix. These matrices include Gaussian kernel function, the minkowski function, the nearest co-relation function and the local scale function. Then, they developed four new algorithms to contrast the clustering results and concluded that building appropriate local scale function is the most available method to make the affinity matrix.

III. PROBLEM FORMULATION

Sumuya Borjigin and Chonghui Guo utilized three stages to determine non-unique cluster numbers of a data set. First of all, they utilized the multiway normalized cut spectral clustering algorithm to make the clusters of the data points of the data set for some cluster number k. Then they used the ratio value of the multiway normalized cut criterion of the obtained clusters and the sum of the leading eigenvalues of stochastic transition matrix as a standard to decide whether the k is a reasonable cluster number. In the third stage, they varied the scaling parameter in the Gaussian function to judge whether the cluster number k is also stable or not.
The algorithms they used in their work are described below:
Algorithm 1: Meil˘*a*-Shi multiway normalized cut spectral clustering algorithm:
– **Input:** Data set $P = \{p1, p2, \ldots, pn\}$, cluster number k.
– **Step 1** Compute the distance matrix W, construct similarity matrix S according to W,
where $W(i, j)$ is the distance between pi and pj , $i = 1, 2, \ldots, n$;
– **Step 2** Compute the Laplacian matrix $L = D - S$;
– **Step 3** Compute the first k eigenvectors $\{v1, \ldots, vk\}$ of the

generalized eigenproblem
$Lv = \lambda Dv$;
– **Step 4** Let $V \in Rn \times k$ be a matrix composed of the vectors $\{v1, \ldots, vk\}$ as columns;
– **Step 5** For $i = 1, \ldots, n$, let $yi \in R1 \times k$ be the vector corresponding to the ith row of V;
– **Step 6** Cluster the points $\{yi \in R1 \times k \mid i = 1, 2, \ldots, n\}$ with the k-means algorithm into
clusters $C1, \ldots, Ck$, if $yi \in Cj$ then $pi \in Pj$, $1 \le i \le n$, $1 \le j \le k$.
–Output: k clusters $P1, \ldots, Pk$.

The similarity matrix used in this algorithm is calculated using Gaussian kernel functions. The Gaussian kernel functions can be divided into global scale Gaussian kernel parameter and the local scale Gaussian kernel parameter.
Global scale Gaussian kernel parameter method is defined as

$$S(i,j)=\exp\left(-\frac{d^2(p_i,p_j)}{\sigma^2}\right)$$

where d(pi,pj) is the distance between point pi and pj and sigma is Gaussian kernel parameter.
Local scale Gaussian kernel parameter is defined as

$$S(i,j)=\exp\left(-\frac{d^2(p_i,p_j)}{\sigma_i\sigma_j}\right)$$

Algorithm 2: Non-Unique Cluster Number determination method based on stability under global scale Gaussian kernel parameter:
– **Input**: Data set $P = \{p1, p2, \ldots, pn\}$, $\epsilon > 0$, $\delta > 0$, user-specified upper threshold
$Cmax \ge 2$ for cluster number to be testified, number of Gaussian kernel parameter $|\Sigma|$.
– **Step 1** Calculate the distance matrix W;
– **Step 2** Let $\sigma min \triangleq \min\{Wi\ j \mid Wi\ j \ne 0, i, j = 1, 2, \ldots, n\}$, $\sigma max \triangleq \max\{Wi\ j \mid i, j = 1, 2, \ldots, n\}$, $\sigma t \triangleq \sigma min + \frac{\sigma max - \sigma min}{\Sigma - 1} * (t - 1)$, for every σt run step 3∼4;
– **Step 3** Calculate the similarity matrix S, where $S(i, j) = \exp\left(-\frac{W_{i,j}^2}{\sigma_r^2}\right)$;
– **Step 4** For every $k = 2, \ldots, Cmax$, make use of the Meil˘a–Shi spectral clustering algorithm
to cluster the data set P into k clusters and calculate the value of index $Ratio(k)$
for obtained clusters;
– **Step 5** To determine whether the candidate cluster number $2 \le k \le Cmax$ is an ϵ-*reasonable* and δ-*stable* cluster number according to the results of step 3 and step 4;
– **Output**: The set of ϵ-*reasonable* and δ-*stable* cluster numbers.

Algorithm 3: Non-Unique Cluster Number determination method based on stability under local scale Gaussian kernel parameter:
– **Input**: Data set $P = \{p1, p2, \ldots, pn\}$, $\epsilon > 0$, $\delta > 0$, user-specified upper threshold
$Cmax \ge 2$ for cluster number to be testified, user-specified maximum number of neighbors
$Kmax \ge 2$.
– **Step 1** Calculate the distance matrix W;
– **Step 2** For $i = 1, 2, \ldots, n$, sort the ith row of W, then calculate $pi\ K$, which is the Kth neighbor of pi , $K = 2, \ldots, Kmax$;
– **Step 3** For $K = 2, \ldots, Kmax$ run step 4∼5;
– **Step 4** Calculate the similarity matrix S, where $S(i, j) = \exp\left(-\frac{W_{i,j}^2}{\sigma_{iK}\sigma_{jK}}\right)$;
– **Step 5** For every $k = 2, \ldots, Cmax$, make use of the Meil˘a–Shi spectral clustering algorithm
to cluster the data set P into k clusters and calculate the value of index $Ratio(k)$
for obtained clusters;
– **Step 6** To determine whether the candidate cluster number $2 \le k \le Cmax$ is an ϵ-*reasonable* and δ-*stable* cluster number according to the results of step 4 and step 5;
– **Output**: The set of ϵ-*reasonable* and δ-*stable* cluster numbers.
The problem that exists in the defined algorithm is of computation cost. The proposed work will be focused to reduce the computation cost and to improve the quality of the clusters.

IV. PROPOSED WORK

In the proposed methodology, the database will be read and segmentation process will be applied and after segmenting the database geometry transformation technique will be applied for asymmetric analysis. When the asymmetric analysis will be done, the relevant and irrelevant data can be clustered and noisy data can be removed and on the basis of data the features will be extracted.

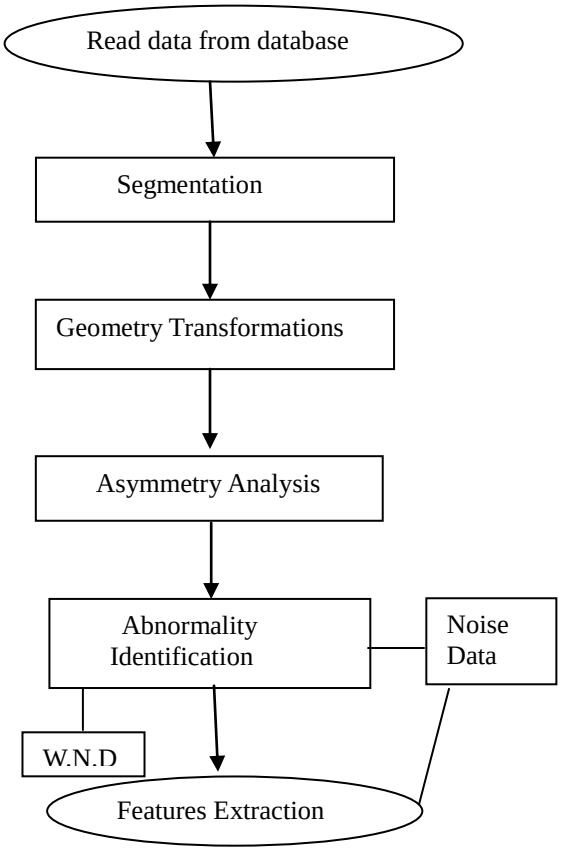

Figure 2: *Flowchart of the Proposed Technique*

V. CONCLUSION

Although the Gaussian kernel functions used for constructing the affinity matrix proved to be a good choice but the computation cost involved using them is quite expensive. The proposed work will be focused on minimising this computation cost for spectral clustering to make this kind of clustering more attractive. Apart from the computation cost, this work will also focus on improving the accuracy and quality of clusters.

REFERENCES

1. Sumuya Borjigin and Chonghui Guo (2012) Non-Unique cluster number determination method based on stability in spectral clustering. Knowl Info System(2013) 36:439-458.
2. http://www.ise.bgu.ac.il/faculty/liorr/hbchap15.pdf
3. Xu-Degang, Zhao Panlei, Gui Weihua, Yang Chunhua, Xie Yongfang (2013) Research on spectral clustering algorithm based on building different affinity matrix.
4. Xianchao Zhang, Quanzeng You (2010), An improved spectral clustering algorithm based on random walk. 5(3):268-278.
5. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
6. http://en.wikipedia.org/wiki/Cluster_analysis
7. http://nlp.stanford.edu/IR-book/html/htmledition/model-based-clustering-1.html
8. http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm
9. Cuimei Guo, Sheng Zheng, Yaocheng Xei and Wei Hao(2010). A Survey on Spectral Clustering.
10. Luxburg U(2007) A tutorial on spectral clustering. Stat Comput 17(4):395-416
11. Hao Huang YunjunGao, Kevin Chiew Lei Chen Qinming He**,** "Towards Effective and Efficient Mining of Arbitrary Shaped Clusters", *ICDE Conference, IEEE 2014*