International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Efficient Way of Determining the Number of Clusters Using Hadoop Architecture

Siri H. P.[1], Shashikala .B[2]

M.Tech. Student Dept. of Computer Science Engineering, BTL Institute of Technology & Management, Bangalore-562125, Karnataka, India

[2]Assistant Prof. Dept. of Computer Science Engineering, BTL Institute of Technology & Management Bangalore-562125, Karnataka, India

Abstract: ***The process of data mining is to extract information from a data set and transform it into an understandable structure. The clustering task plays a very important role in many areas such as exploratory data analysis, pattern recognition, computer vision, and information retrieval. The key idea is to view clustering as a supervised classification problem, in which we estimate the "true" class labels. The problem of determining the valid number of clusters is not easy. To overcome this problem many well known methods are used to find a correct number of clusters i.e. Gap statistic, Path based clustering and Figure of Merit (FOM) but these methods could not solve the problem of finding number of clusters efficiently. This paper focuses on "Average Intracluster Distance" index to validate the estimated number of arbitrary shaped clusters. In hadoop the proposed technique is based on the local relations between patterns and their clustering labels which makes use of Minimum Spanning Tree (MST) algorithm based on the multiplicity property of MST to get accurate results in efficient manner .***

Keywords: Minimum Spanning Tree (MST), Gap statistic, IC-av.

1. Introduction

Cluster analysis is an important and widely used for determining the clusters of known and unknown class labels using clustering methods. Clustering is the process of orchestrating data into meaningful groups, and these groups are called clusters. It identifies groups of related records that can be used as a starting point for exploring further relationships. Clustering can be seen as a generalization of classification. Classification is more similar to just finding "where to put the new object in". Clustering on the other hand analyzes the data and finds out characteristics in it, either based on responses (supervised) or more generally without responses (unsupervised). The proposed methodology introduces a new approach for determining number of clusters. To achieve this, the new approach uses minimum spanning tree for the given dataset.

The most local relationship between the datapoints is one of the first neighbours and the connected graph that better follows this concept is the minimum spanning tree. The main part of work done on this paper is a new clustering validation index based on a new distance called average intracluster distance measure that can find unsupervised clusters.

This paper describes the previous work done on clustering validation, clustering validation methods also a cluster validation index called average intracluster distance. The results of the validation index are obtained parallel in the hadoop environment.

2. Related Work

The validation of the clustering process is based on the comprehension of biological information related to the problem. The use of this kind of information helps to define "natural groups," and thus, it helps to find meaningful clusters in problems where that knowledge is valid. However, this type of methods can be too focused on the problem at hand, which can finally lead to an analysis that is only valid for a unique set of data.

The work by Yeung et al., gives a clear example of Graphical validation methods that rely on different working principles but the nature of their output demands that the analyst selects a solution. The authors proposed Figure of Merit (FOM), a method that assumes that clusters can be validated by an unseen condition. To study the likeliness of a c-cluster model solution, the authors analyze the average response on unobserved features. The unobserved features are used in a validation step to provide an estimate of the error of the clustering solution on these conditions. As output, the method draws a response curve, from which the user must select the most appropriate number of clusters using a different working principle [1].

A. Ben-Hur proposed "natural grouping" to describe patterns forming clusters described by a hidden definition, i.e., clusters formed by a rule unknown to the analyst. Once the "natural groups" are defined, it is possible to use the previous similarity function or clustering validation index to find the most likely number of clusters for a given clustering algorithm. The work by I. Dhillon introduces a new validation measure that can detect arbitrary-shaped clusters. To achieve this, the new cluster validation index uses the minimum spanning tree (MST) of the data. The only assumption made about clusters, to guarantee their detection, is that the maximum first-neighbor distance should be less than clusters separation [2].

Monti et al. proposed a method named Consensus Clustering. The validation process is based on many clustering rounds, i.e., clustering ensembles, which allows the user to visualize the different clustering solutions to inspect its constancy. This method uses a robotic ad hoc rule based on the difference in the area between successive cumulative density function defined by the clustering solution on the consensus

matrix. The graphical output based on this rule is used to select the number of clusters. The authors demonstrated that their results are corresponding to the ones of the gap statistic [3]. However, the results of Monti et al. can be difficult to discriminate since the solution may not be unique. To analyze the reasons of multiple clustering solutions, one would need to analyze consensus matrices or the cumulative density functions. In general, graphical output requires proficient knowledge to understand results and find an answer, even in the cases of automatic methods like Prediction Strength and Consensus Clustering.

R. Tibshirani developed a gap statistic method. This method uses Within Cluster Sum of Squares (WSS). A kernel gap statistic method would require to change the within clusters sum of squares by a kernel to detect nonglobular clusters [4].

Tibshirani and Walther, developed a method for clustering validation called Prediction Strength that is based on applying two rounds of clustering, first to a training set and then to a test set. The resulting train and test labels are used to create a co-membership matrix. This is used by the authors to outsmart the problem of labels assignment between train and test sets. The prediction strength index is then calculated over the co-membership matrix. This measure represents the property of the hypothesis that the data have c clusters [5].

Dhillon et al., proposed an interesting weighted kernel kmeans algorithm. The optimization rule that the authors describe for kernel k-means could be also used for a Kernel gap statistic method. Merging the optimization rule from Dhillon et al. with the gap statistic could be useful to find nonglobular clusters, although this modification comes with the challenge to set extra parameters. The key feature of these methods is that they have a simple automatic rule that informs its user the number of clusters detected. This quality is most important for users with little or no clustering knowledge. This paper searches the use of an MST to find clusters with arbitrary shapes [6].

A. Azaran developed a "Spectral Methods for Automatic Multiscale Clustering". The method is related to changing scales or clusters densities that limit the power of the Gaussian Kernel to describe the data. This, in turn, leads Spectral Clustering to bad clustering solutions. For these cases, kernel gap may also experience problems in combination with a Gaussian Kernel because of the direct relation to Spectral Clustering [7].

Pihur et al. developed an automatic method that was making use of a set of validation indexes to place a group of clustering algorithms for a given clustering task. This method automatically selects the best clustering algorithm by simultaneously testing multiple validation methods [8].

The validation of the clustering process is based on the comprehension of biological information related to the problem. The use of this kind of information helps to define "natural groups," and thus, it helps to find meaningful clusters in problems where that knowledge is valid. However, this type of methods can be too focused on the problem at hand, which can finally lead to an analysis that is only valid for a specific set of data [9].

In the validation literature, there are many boulevard of research that aim to solve relevant validation problems. The present work has a narrower focus, aiming at solving some inadequacy of previous validation methods. We limit ourselves to the comparison of single validation methods.

Ariel E. Baya and Pablo M. Granitto proposed an average Intracluster Distance (IC-av). The IC-av represents the average sum of the maximum edges between all pairs defined along the Minimum Spanning Tree (MST) algorithm. To evaluate the closeness between the clusters the IC-av makes use of local relationship between the points. Thus, it can be used to detect the correct number of clusters [10].

The main contribution of this work is a new clustering validation index based on IC-av distance measure that can find number of unsupervised clusters parallel in the hadoop environment.

3. Existing System

This section describes some of the validation methods that are used to validate the quality of the solution given by the previous methods that are used to find the correct number of clusters for a given data set.

A. Consensus Clustering

Monti et al. [5] developed a method named consensus clustering. This validation process is based on many clustering rounds, i.e., clustering collections, which allows the user to visualize the different clustering solutions to inspect its stability. This method uses an automatic ad hoc rule based on difference in the area between successive cumulative density function defined by the clustering solution on the consensus matrix. The graphical output based on this rule is used to select the number of clusters. The author showed that their results are comparable to the ones of the gap statistic [6]. The results of Monti et al. can be difficult to discriminate since the solution may not be specific. To analyze the reasons of multiple clustering solutions, one would need to analyze consensus matrices or the cumulative density functions. In general, graphical output requires proficient knowledge to understand the results and find an answer, even in the cases of automatic methods like Prediction Strength and Consensus Clustering. For an instance consider a cluster ensemble C = {c1, c2, cn} of n data points X = {x1, x2, xn} as an input to clustering algorithm. Compute consensus matrix using a set of connectivity matrices. The consensus matrix M is an n*n matrix such that $M(C)_{ij}$ = number of times object x_i in the ensemble C. for a given clustering C_i, define an adjacency matrix A_i as,

$$Aij = \begin{cases} 1, \ if \ object \ xi \ was \ clustered \ with \ xj \\ 0, \ otherwise \end{cases}$$

Thus, consensus matrix M would be the sum of adjacency matrix of each clustering in the collection $M(C) = \sum_{i=1}^{N} Ai$. The entry (i, j) in the consensus matrix records the number of times i and j are assigned to the same cluster divided by the total number of times as selected. It should be clear that consensus matrix is symmetric i.e. M(i, j) = M(j, i) for all i and j. example : C1 = { {1, 2, 3, 4} {5, 6, 7, 8, 9} {10, 11}} , k1=3. C2 = {{1, 2, 3, 4} {5, 7} {6, 8, 9} {10, 11}}, K2=4.

C3 = {{1, 2} {3, 4} {5, 6, 8} {7, 9} {10, 11}}, k3=5. The consensus matrix for the ensemble is as follows:

	1	2	3	4	5	6	7	8	9	10	11
1	3	3	2	2	0	0	0	0	0	0	0
2	3	3	2	2	0	0	0	0	0	0	0
3	2	2	3	3	0	0	0	0	0	0	0
4	2	2	3	3	0	0	0	0	0	0	0
5	0	0	0	0	3	2	2	2	1	0	0
6	0	0	0	0	2	3	1	3	2	0	0
7	0	0	0	0	2	1	3	1	2	0	0
8	0	0	0	0	2	3	1	3	2	0	0
9	0	0	0	0	1	2	2	2	3	0	0
10	0	0	0	0	0	0	0	0	0	3	3
11	0	0	0	0	0	0	0	0	0	3	3

Figure 3.1: The Consensus Matrix for Sample Data

In the given consensus matrix, each row and column headers represent the datapoints in the clusters C1, C2, and C3. For the cluster C1 it has three different groups of datapoints, also similarly in C2 it has four and in C3 it has two different groups of datapoints. The values inside the matrix are written based on the number of times the object x_i is clustered with the x_j. In the above consensus matrix, each cluster representing a block diagonal value as 3 in the red rectangular box. Hence the resulting consensus matrix is clearly indicates with $k^*= 3$. But for the high dimensional datasets the interpretation of consensus matrix is very complex to determine the valid number of clusters [3].

B. Gap Statistic

The gap statistic is a method for estimating the optimal number of clusters. This technique is based on the idea that the change in within-cluster dispersion with the increase of the number of clusters and is expected under a reference distribution for random data. First, assume that there are a set of samples $\{x_i\}$ then by use of the clustering method, the resultant clusters *C*1, *C*2 . . . *Ck* can be obtained. For any cluster *Cr*, the sum of the pair wise distances d^2 $(x_i, x_{i'})$, for all points in cluster *r* is calculated. And the sum of within-cluster dispersion W_k is defined as the following equation

$$WSS = \sum_{r=1}^{c} \frac{1}{Nr} \sum_{i,k \in Lr} d^2(x_i, x_k), Eqn.(3.1)$$

Where d = The Euclidean distance and n_r = the number of clusters. K= 1, 2, K clusters. Euclidean distance is used to calculate the distance between the i^{th} observation and the k^{th} cluster. The index presented in the eqn. (3.1) estimates cluster closeness. The gap statistic uses the kmeans algorithm to estimate the optimal number of clusters for the given dataset. For an instance consider iris dataset, the data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris Virginica and Iris Versicolor). Four features were measured for each sample: the length and the width of the sepals and petals, in centimeters. The result of Within Cluster Sum of Square (WSS) for the iris dataset is as follows:

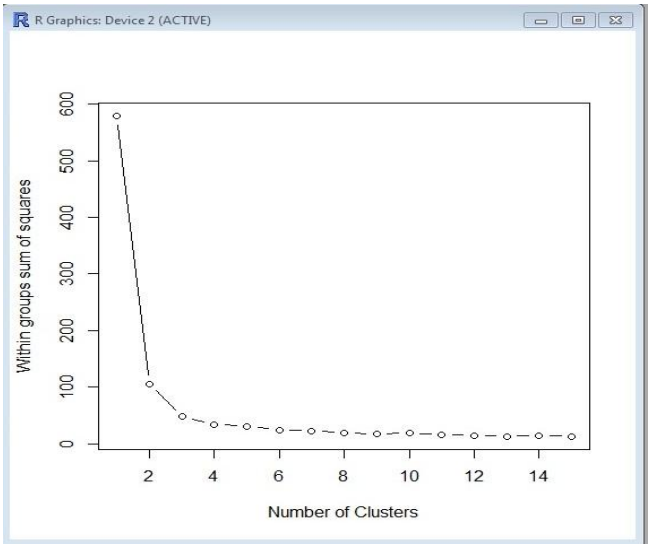

Figure 3.2: The WSS of Iris Dataset.

The Fig. 3.2 plot is the within cluster sum of square distance versus the number of clusters. It indicates that there is a distinct drop in within group's sum of squares when moving from 1 to 3 clusters. After three clusters, this decrease drops off, suggesting that a 3-cluster solution may be a good fit to the data. The estimated number of cluster for the iris data set is 3. A curve in the graph can suggest the appropriate number of clusters. We can see in the above graph the WSS from third cluster it is sequentially plotted as though number of cluster increases. It means that the valid number of clusters for the iris dataset is three. In contrast to the result of WSS, the gap statistic for the iris dataset is as follows:

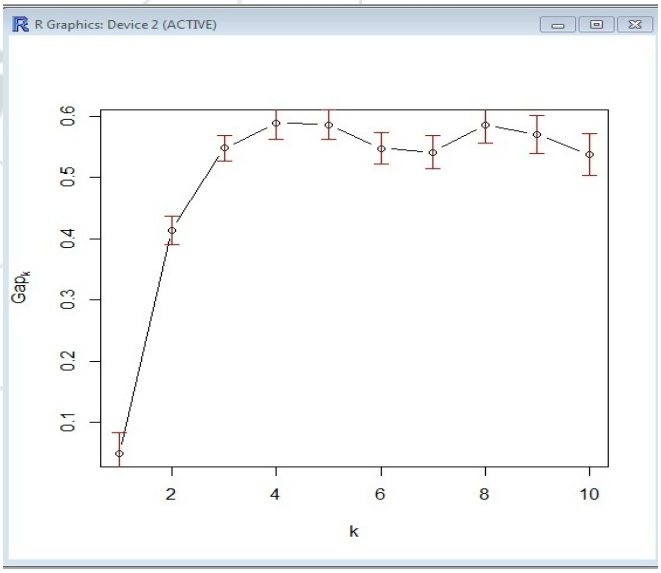

Figure 3.3: The Gap Statistic for Iris Dataset

The figure 3.3 plot is the gap versus the number of clusters k. From the above gap statistic plot it is very difficult to determine the optimal number of clusters in the case of high dimensional dataset because the results of the gap statistic are not perfectly discriminated. Hence to overcome these problems, the use of average intracluster distance measure is processed in the Hadoop environment to estimate the valid number of clusters in a well suitable form.

4. Proposed System

This section describes an approach for determination of cluster using the hadoop environment. If possible multiplicity of MST property is true, and then the MSTs can be processed parallel in the Hadoop environment in order to get more accurate results in an efficient manner. This section describes the algorithms like Minimum Spanning Tree (MST) which is used to deduce a pairwise distance to discover non-spherical groups of clusters more accurately. The concept of average Intracluster distance method is used to find the valid index for determining the correct number of cluster for a given data set.

A. Determination Of Cluster In The Hadoop Environment

Hadoop is a powerful open-source platform for handling massive datasets at scale by processing the big data in a distributed fashion on large clusters of commodity hardware. The use of hadoop in the cluster analysis is designed specifically for storing and analyzing huge amount of unstructured data in a distributed computing environment. In the proposed methodology, the dataset is given as an input to clustering process model. For the given dataset an undirected connected minimum spanning tree (MST) graph is generated. If the possible multiplicity of MST property holds true, then for each MST graph the pairwise maximum edge distance (MED) is calculated. Using the MED the average Intracluster index is obtained to determine the cluster solution in the hadoop environment. The NameNode in the hadoop environment stores the location of the given data. The DataNode process the dMED and IC-av functions parallel in the hadoop environment which produces the results in an efficient timely manner.

1) Partitioning of MST Based on MED: The partitioning of minimum spanning tree organizes the objects into several exclusive groups of cluster in order to define the average Intracluster index. Given an undirected, connected, and weighted MST graph, based on the pairwise maximum edge distance, the graph is partitioned into two groups. Again it is repeated recursively until it confines to a single cluster. The result of each partitioned MST is used to define the average Intracluster distance for each of the MSTs.

2) The Average Intracluster Distance: The Average Intracluster Distance method used to detect the correct number of clusters in many arbitrary shaped and globular clustering shapes. The gap statistic method is combined with average Intracluster (IC-av) index to detect number of arbitrary shaped clusters present in the given data set. The set $X \varepsilon R^d$ of N data points $\{X_1, X_2 \ldots X_N\}$ where each point x_i is described by d features x_{ij} (j = 1, 2, d). The set L is a clustering solution, a set of labels that divides X in c clusters {L1, L2 . . . L_c}. Both sets X and L can be used as parameters of a function that measures the quality of the clustering solution L in X. A common example of this is the WSS. The index presented in the Eqn. (3.2.1) estimates cluster closeness, but instead of assuming spherical shape, it assumes that clusters are connected structures with arbitrary shape. Patterns from data set X are used to construct an undirected complete graph G (V, E) = $x_{i \equiv} v_i$ and edge e_{ik} corresponds to the pairwise distance between vertex v_i and v_k.

The idea followed by the new validation index is to use local relationship between points to evaluate the global closeness between the clusters formed in X. The most local relationship between points is the one of first neighbours and the connected graph that better follows this concept is the minimum spanning tree. A MST (V, E^t) where $E^t \subset E_r$, uses local information to form a graph that joins all vertex. This structure restricts the original Euclidean space to the paths formed by the edges of the MST. From the set of edges E^t, it is possible to deduce a pairwise distance to detect non spherical groups of clusters more accurately. In this work, the pairwise distance between vertexes is defined by the longest edge in the path joining a pair of points. Validating, $P_{ik}=G'(V_p, E_p{}^t)$ is a sub graph from MST (V, E^t) where ($V_p \subset V$) and ($E^t_p \subset E^t$) are the subsets of vertex and edges that form the path between vertex v_i and v_k where $v_{i,} \epsilon v_k$ and v_k are part of the path. Using the previous notation, a new distance named Maximum Edge Distance (MED) is defined as

$$d^{MED}(v_i, v_k) = d_{ik}^{MED} = \left\{ E_p^t \in \frac{P_{ik}}{\max(e_p^t)} \right\}, \text{(Eqn 4.1)}$$

In the Eqn. (4.1), P_{ik} represents the longest edge path. e^t_p represents the maximum weighted edge in the MST path, $v_{i,}v_k$ are the vertices of single edge in the MST graph. E^t_p is the subset of vertex and edges that from the path between vertex vi and $v_{k.}$ The intracluster closeness among the members of a partition can be defined as the average of the pairwise MED distance among those members. This leads to the definition of the average intracluster gap:

$$IC - av = \sum_{r=1}^{c} \frac{1}{n_r} \sum_{i,k \in L_r} d^2_{MED}(x_i, x_k) \text{ , (Eqn 4.2)}$$

Where in the Equations (4.2), C – clusters, r - represents class label, n_r is the number of cluster L_r and d (...) is the pairwise maximum edge distance and $x_{i,}x_k$ are the values of the datapoints. The Equation (4.2) and (3.1) are the same: they only dissent in the averaged metric. While the WSS has a simple interpretation, it is the squared distance toward the clusters center, the IC-av represents the average sum of the maximum edges between all pairs defined along the minimum spanning tree. This measure can be interpreted as an idea of the closeness of the points in each cluster. It is a measure that considers local relationships and that makes no assumptions about clusters shape. Thus, it can be used to detect the correct number of clusters in many arbitrary-shaped and globular clustering shapes. To actually detect the number of clusters, the new metric is combined with the gap statistic. This method was shown useful by the authors to detect spherical shaped clusters. The use of the new IC-av will amend the previous method so it can detect groups of patterns with different shapes.

B. System Architecture

The architecture design process is concerned with establishing a basic structural framework for a system. It involves identifying the major components of the system and communications between these components.

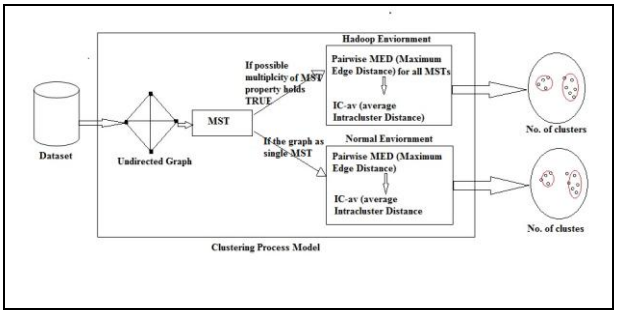

Figure 4.2.1: Architecture of Proposed Methodology

C. Modules

1) Construct an Undirected Complete Graph for the Given Dataset: A graph is a representation of a set of objects where some pairs of objects are connected by links. The interconnected objects are represented by mathematical abstractions called vertices, and the links that connect some pairs of vertices are called edges. An undirected graph is one in which edges have no orientation. The edge (a, b) is identical to the edge (b, a), i.e., they are not ordered pairs, but sets {*u*, *v*} (or 2-multisets) of vertices. The maximum number of edges in an undirected graph without a self-loop is $n(n - 1)/2$.

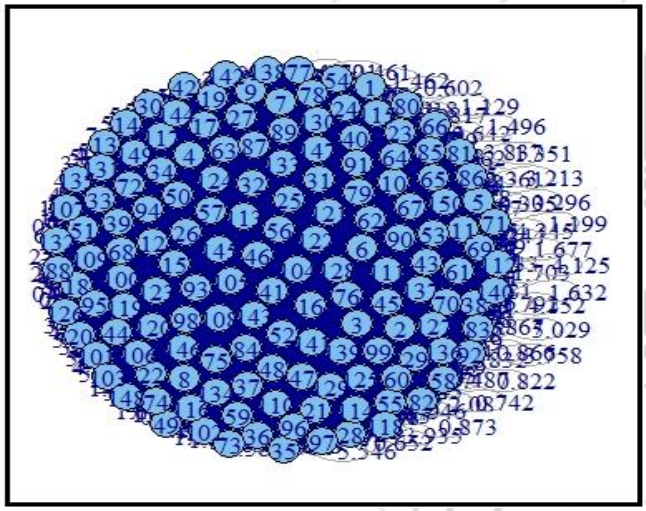

Figure 4.2.2: A simple Undirected Complete Graph of Iris dataset.

2) Create a Minimum Spanning Tree (MST) for the Graph : Given a connected, undirected graph of a dataset, a spanning tree of that graph is a subgraph that is a tree and connects all the vertices together. A single graph can have many different spanning trees. We can also assign a weight to each edge, which is a number representing how unfavorable it is, and use this to assign a weight to a spanning tree by calculating the sum of the weights of the edges in that spanning tree. The most local relation between datapoints is the one of first neighbours and the connected graph that better follows this concept is the minimum spanning tree. A MST (V, E^t), where $E^t \subset E$, uses local information to form a graph that joins all vertex. This structure limits the original euclidean space to the paths formed by the edges of the MST.

To find minimum spanning tree for the graph it uses Prim's algorithm. **Prim's algorithm** is a greedy algorithm that finds a minimum spanning tree for a connected weighted undirected graph. This means it determines a subset of the edges that forms a tree that includes every vertex, where the total weight of all the edges in the tree is minimized.

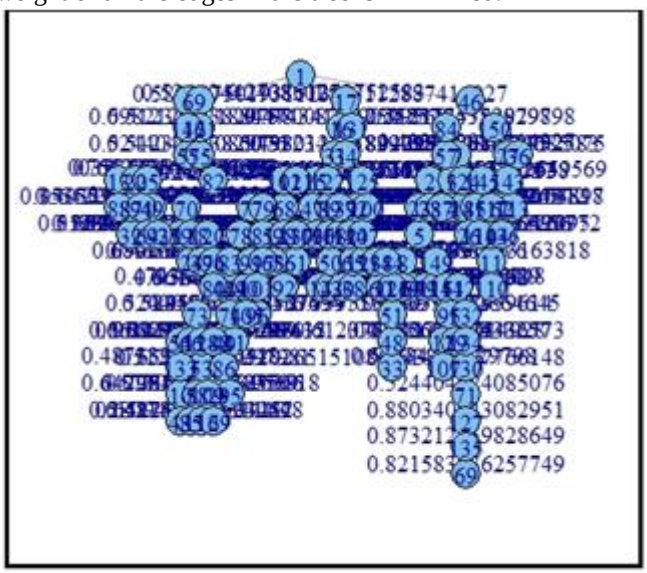

Figure 4.2.3: MST of an Undirected Graph for Iris dataset

3) Partition the Dataset and Compute Average Intracluster Distance: The intracluster closeness among the members of a partition can be defined as the average of the pairwise MED distance among those datapoints. From the set of MST edges E^t, it is possible to deduce a pairwise distance to detect nonspherical groups of clusters more accurately. In this work, the pairwise distance between vertexes is defined by the longest edge in the path joining a pair of points. Validating, $P_{ik}=G(V_p,E_p^t)$ is a subgraph from $MST(V, E^t)$ where $V_p \subset V$ and $E^t_p \subset E^t$ are the subsets of vertex and edges that form the path between vertex vi and v_k where v_i, $v_k \in V_p$, i.e., v_i and v_k are part of the path. Using the previous notation, a new distance named maximum edge distance (MED). The distance is symmetric, $d_{ik}^{MED} = d_{ki}^{MED}$; always positive $d_{ik}^{MED} >= 0$; satisfies identity, $d_{ii}^{MED} = 0$ and also the triangle inequality.

The IC-av represents the average sum of the maximum edges between all pairs defined along the MST (Minimum Spanning Tree). This measure can be interpreted as an estimation of the closeness of the points in each cluster. It is a measure that considers local relations and that makes no premises about clusters shape. Thus, it can be used to detect the correct number of clusters in many arbitrary-shaped and globular clustering shapes. To actually detect the number of clusters, the new metric is combined with the gap statistic. In this method, the number of clusters c increases from 1 to c_{max}, where c_{max} is greater than c* ($c_{max} > c^*$) and c* represents the true number of clusters that the method is trying to detect. The average sum given by (3) measures the changes in the clustering solutions as the number of partitions c varies from 1 to c_{max}.

5. Conclusion

The proposed methodology focuses on the computation of intracluster distance in the hadoop environment. This is achieved by constructing the graph for the given dataset. Using the undirected graph, the MST (Minimum Spanning Tree) of the graph is constructed. This MST graph is taken

as an input to obtain the pairwise Maximum Edge Distance (MED).MED is used to compute the average Intracluster Distance. This method is based on the relationship between patterns and their clustering labels. The dMED and IC-av functions can be processed parallel in the hadoop environment in order to produce more accurate results in an efficient timely manner for both artificial and gene expression dataset.

References

[1] K.Y. Yeung, D.R. Haynor, and W.L. Ruzzo, "*Validating Clustering for Gene Expression Data*," Bioinformatics, vol. 17, no. 4, pp. 309-301, 2001.
[2] A. Ben-Hur and A. Elisseeff, and I. Guyon, "*A Stability Based Method for Discovering Structure in Clustered Data,*" Proc. Pacific Symp. Biocomputing, pp. 6-17, 2002.
[3] S. Monti, P. Tamayo, J.P. Mesirov, and T.R. Golub, "*Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data*," Machine Learning, vol. 52, nos. 1/2, pp. 91-118, 2003.
[4] R. Tibshirani, G. Walther, and T. Hastie, "*Estimating the Number of Clusters in a Data Set via the Gap Statistic*," J. Royal Statistical Soc. B, vol. 63, pp. 411-423, 2003.
[5] I. Dhillon, Y. Guan, and B. Kulis, "*Kernel K-Means, Spectral Clustering and Normalized Cuts*," Proc. Int'l Conf. Knowledge Discovery and Data Mining, pp. 551-556, 2004.
[6] R. Tibshirani and G. Walther, "*Cluster Validation by Prediction Strength*," J. Computational & Graphical Statistics, vol. 14, no. 3, pp. 511-528, 2005.
[7] A. Azran and Z. Ghahramani, "*Spectral Methods for Automatic Multiscale Data Clustering*," Proc. Conf. Computer Vision and Pattern Recognition, pp. 190-197, 2006.
[8] V. Pihur, S. Datta, and S. Datta, "*Weighted Rank Aggregation of Cluster Validation Measures: A Monte Carlo Cross-Entropy Approach*," Bioinformatics, vol. 23, no. 13, pp. 1607-1615, 2007.
[9] Z.V. Volkovich, Z. Barzily, G.-W. Weber, and D. Toledano-Kitai, "*Cluster Stability Estimation Based on a Minimal Spanning Trees Approach*," Proc. AIP Conf., pp. 299-305, 2009.
[10] G. Stegmayer, D.H. Milone, L. Kamenetzky, M.G. Lpez, and F. Carrari, "*A Biologically Inspired Validity Measure for Comparison of Clustering Methods over Metabolic Data Sets*," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 9, no. 3, pp. 706-716, May/June 2012.
[11] A.E. Baya and P.M. Granitto, "*How Many Clusters: A Validation Index for Arbitrary Shaped Clusters*," BMC Bioinformatics, vol. 10, article 2, April 2013.