ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

Clustering Optimal Algorithm- A Survey

B Kalai Selvi[1], M Ashwin[2]

PG Scholar, Department of CSE, Adhiyamaan College of Engineering, Hosur, Tamil Nadu, India[1]

Associate Professor, Department of CSE, Adhiyamaan College of Engineering, Hosur, Tamil Nadu, India[2]

A**BSTRACT***:* Data Mining is the process of extracting the information from the raw data. One of the challenging areas in the data mining is the clustering and finding the best cluster is difficult task. The aim of this paper is to find a clustering algorithm which fits with consensus clustering. Consensus clustering is the process of finding the best clustering from multiple clustering. The problem in consensus clustering is that it doesn't fit well for all clustering algorithm. This paper is to find the best clustering algorithm and it is embedded with consensus clustering. Here we focus on 3 different algorithms such as k-means algorithm, hierarchical algorithm and graph partitioning algorithm. The main focus of this paper is to find the algorithm which fits for large datasets. We summarize our observation and identify the common pitfalls in the previous works.

KEYWORDS: Clustering, Consensus Clustering, Data Mining, k-Means, Hierarchical

I. INTRODUCTION

The growth of data such as text, images, music, videos and other data's are increased in the past decades. All those data's are considered as raw data. To obtain valid information from that raw data the user has to mine the information. To extract the valid information from raw data we use an excellent technique called data mining. Data mining is the process of extracting information or discovering knowledge from large amount of raw data [1]. The steps involved in data mining are cleaning, integration, selecting, transformation, pattern evaluation. The data mining consists of wide variety of concepts to get the knowledge.

The focus of the paper is only on the clustering techniques. Clustering is the process of grouping the data according to some criteria that most similar data to form a cluster. The groups are not predefined in clustering. K-means clustering, Hierarchical clustering, Graph based partitioning, fuzzy c-mean clustering are some of the commonly used clustering methods. Basically, clustering technique is divided into soft cluster and hard cluster. In hard cluster the data must present only in one cluster whereas in soft clustering the data may present in more than one cluster. K-means clustering algorithm is mainly used in hard clustering and fuzzy clustering algorithm is used in soft clustering.

Consensus clustering is also called as aggregation of clustering or cluster ensemble [2]. Consensus means optimal and cluster is group i.e., selecting a single cluster which may fit better than other clustering. To a particular datasets the different clustering algorithms are used and clusters are obtained .The aggregation of clusters is sent to the consensus clustering to find the cluster which fits for the problem.

The three different clustering algorithms such as k-means based clustering; hierarchical clustering and graph partitioning are focused in this paper. The k-means clustering partitions the dataset into k clusters according to user's choice. In Hierarchical clustering the similarity between two data's are identified and grouped. In graph partitioning the data is represented in the form of graph and partitions are done according to some properties.

The paper is organized as follows: Section II contains the related work and Section III explains about the evaluation methodology of clusters. Section IV indicates the observation and Section V discusses about the sample datasets of the methodologies. Finally, Section VI proposes some suggestions and conclusion.

II. RELATED WORK

In the past decades data mining is growing rapidly. One of the main parts of data mining is clustering. Clustering is used in many day-to-day applications such as placing telephone booth / towers, opening a hospital. In this

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

section we review the clustering techniques and the use of clustering. The clustering techniques that are reviewing are k-means, hierarchical clustering and graph partitioning and how the clustering is embedded with consensus clustering.

Bishnu and Bhattacherjee [3] say about using quad tree algorithm with k-mean. In this quad tree algorithm each dataset is divided into 4 equal parts, similarly the subpart are also divided according to the clustering measures. The quad-tree is combined to k-means to initialize the cluster size .The quad tree based k-means is used to find the faults in the programming module. Similarly, Xiaojun Chen et al. [4] proposed a paper with k-means algorithm embedded with two-level variable weighting clustering algorithm (TW) which is used to obtain the weights of views and individual variables. The TW algorithm has two levels such as important variable is selected in first level whereas in second level k-mean algorithm is used find the cluster structure to find the clusters.

The hierarchical clustering is classified into agglomerative and divisive .In agglomerative hierarchical clustering each individual data is clustered according to the closest distance where as in divisive clustering the data set is divided into individual data. Shoa et al. [6] proposed that mean shift is used to speed up the agglomerative hierarchical clustering to check the performance of the data. Guangxia Li et al. [7] says that graph based method is used to improve the transductive support vector machines (SVM), which is an existing semi-supervised learning algorithm.

To improve the quality of individual data the methodology such as Cluster-based similarity partitioning (CSPA), Hyper graph partitioning Algorithm (HGPA), and Meta- clustering Algorithm (MCLA) is used. In CSPA similarity between two data points are defined and placed in the same cluster. In HGPA all clusters are equally weighted and in MCLA clusters are again clustered to solve the particular problem [10].

III. EVALUATION METHODOLOGY

We conducted a survey of research work in the area of clustering. Here we considered three different clustering algorithms such as k-means clustering, Hierarchical clustering and graph partitions that are explained below:

K-MEANS CLUSTERING

The aim of the k-means clustering is to partition the data into k-clusters according to the center. The data which are near to the center is formed a cluster. The general equation of K-means based clustering is as follows:

$$V = \sum_{i=1}^{C} \sum_{j=1}^{C_i} ||x - y||^2$$

Where, V is k-means clustering, C is the number of cluster centers, C_i is the data points in the i^{th} cluster, $||x–y||^2$ is the squared Euclidean distance.The k-means algorithm can use other distance finding algorithm such as Kullback - Leibler divergence also called as information divergence or KL divergence, cosine distance and L_p distance. But mostly used distance algorithm is the Euclidean distance or Squared Euclidean distance.

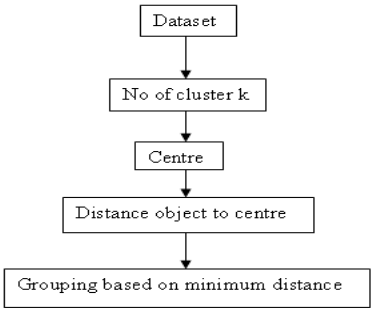

Figure 1: k-means Clustering Steps

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

The distance formula is used to find the distance between the center point and other data points from that clusters are made which has smallest distance. The k-means clustering is used in many places like placing towers, location of railway station, conducting world cup games, starting a shop. Figure 1 describes about the working flow of the k-means algorithm.

Input: D: Data set
Output: C: Cluster
Procedure:

1. Get the dataset as input
2. Cluster center and number of clusters is initialized
3. Compute the distance between the data center C_i and all data using squared Euclidean distance$||x-y||^2$
4. Gather the data to form the cluster C which are close to data centers

AGGLOMERATIVE HIERARCHICAL CLUSTERING

The hierarchical clustering builds the tree like structure to form a cluster. This hierarchical clustering is of two types such as divisive (top-down approach) and agglomerative (bottom-up approach).Here we mainly focus on the Agglomerative clustering where the individual data's are combined to form a cluster in hierarchical manner. The hierarchical clustering is used in retrieving documents i.e., it is mainly used in information retrieval and also used while searching in net. Figure 2 describes about the work flow of the agglomerative clustering.

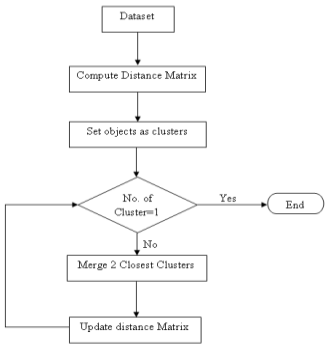

Figure 2: Agglomerative Hierarchical Clustering Steps

Input: D: Data Points in distance matrix
Output: C: Cluster
Procedure:

1. Initially make level L(0) as 0 and sequence number m=0
2. The least distance for the pair of clusters are found using d[(r),(s)]=min dis [(i),(j)] for overall pairs
3. Sequence number is increased accordingly as m=m+1 and merging clusters are indicated as r,s and level L(m)=d[(r),(s)]
4. Update distance matrix ,D D[(k),(r, s)]=min(d[(k),(r)],d[(k),(s)])
5. The above steps is repeated until it forms a single cluster

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

GRAPH PARTITIONING ALGORITHM

In graph partitioning algorithm the data's is represented in the form of graph with edge E and vertex V. The k-way partition is used to divide the vertices to form small components. Some of the example for graph partitioning is Kernighan-Lin Graph Partitioning Algorithm, Spectral Graph Partitioning.

CONSENSUS CLUSTERING

Consensus Clustering is also called as aggregation of clustering or cluster ensemble. Consensus clustering is used to find the optimize solution for a problem. Different clustering algorithms are combined with consensus clustering to form the clusters. Consensus clustering fits for both hard and soft clusters. The consensus clustering has three schema/ algorithm to find the particular optimal solution for the problem. The algorithms are Cluster-Based Similarity Partitioning Algorithm (CSPA), Hyper-Graph Partitioning Algorithm (HGPA), Meta Clustering Algorithm (MCLA).In CSPA the similarity between two data points are defined and group the data which has higher chance to form a cluster. In HGPA the cluster data is repartitioned and clusters are equally weighted. In MCLA the clusters are clustered and the clusters are sorted to the corresponding problem. Some examples for consensus clustering are categorical data, mainly for privacy preserving scenarios. The general equation of consensus clustering is defined below:

$$CC(\pi, S) = \sum_{i=1}^{n} W_i \cup (\pi, \pi_i)$$

Where, CC is Consensus Clustering, π is the cluster, S is the set of clusters such as S=$\{\pi_1, \pi_2, \pi_3, \ldots\ldots, \pi_n\}$, W is the weight of the cluster, n is the partition .Figure 3 explains about the workflow of the consensus clustering.

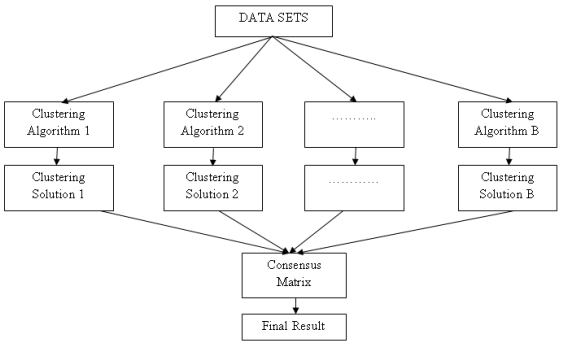

Figure 3: Consensus Clustering Steps

Input:	D, Data Points in distance matrix Clustering algorithm K, Set of clusters h, Sampling Iterations
Output:	P, Partition

Procedure:

1. Initially make connectivity matrix M = 0
2. $D^{(h)}$ = Resample(D)
3. $M^{(h)}$ = Cluster(D,K)
4. M = M∪$M^{(h)}$
5. Compute Consensus Matrix M^K
6. Partition D into k clusters using M^K

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

IV. OBSERVATION

TABLE 1
Analyzed Papers and Results

TITLE	METHODOLOGY	PROS	CONS
Software Fault Prediction Using Quad Tree-Based K-Means (QDK) Clustering Algorithm [3]	Quad Tree based Algorithm is applied for finding initial cluster centers for K-Means algorithm	1. QDK cluster quality is high when compared to k-means clustering 2. When compared to Catal et al. Two stage approach (CT) [1], Catal et al. Single stage approach(CS),QDK performance, effectiveness is better	1. K-means has less iteration when compared to QDK
TW-k-Means: Automated Two-Level Variable Weighting Clustering Algorithm for Multiview Data[4]	Two-level variable weighted clustering algorithm with k-means	1. Execution time of TW-k-means is more when compared to normal clustering algorithms like locally adaptive clustering (LAC), weighted combination of exemplar-based mixture models (WCMM) 2. Produces best result for multiview datasets	It does not show the best result for individual variable data set
Agglomerative Mean-Shift Clustering[5]	Agglomerative hierarchical clustering	1. Reduces the costly MS clustering while using agglomerative clustering 2. Agglomerative Mean Shift clustering is fast, stable and accurate	The size of dataset must be less than or equal to 500 instances
Hierarchical Co-Clustering: A New Way to Organize the Music Data[9]	Divisive hierarchical clustering Agglomerative hierarchical clustering	These two methods have a capability to achieve better performance	Take time to do as dividing and combining process takes place
Synchronization-Inspired Partitioning and Hierarchical Clustering[6]	Minimum Description Length (MDL) principle, it allows partitioning and hierarchical clustering	Works well for small database	Does not fit for large data sets
Multiview Semi-Supervised Learning with Consensus[7]	Graph-Based Methods	1. 10 percent better than a single-view learning approach, 2. Accurate and stable	It does nt produce stable result for single-view learning
Identifying the Most Connected Vertices in Hidden Bipartite Graphs Using Group Testing[12]	MVC algorithm	Determines the size of the vertices to be input to each group	Linear testing is costly

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

Mining Graph Topological Patterns: Finding Covariations among Vertex Descriptors[11]	Top Graph Miner(to discover sensible Patterns)	Frequently used attribute and pattern is obtained	Small data's takes large time
Consensus-based ensembles of soft Clusterings [8]	Consensus clustering (soft clustering-Graph partitioning, fuzzy c-means, Hard clustering-CSPA,HGPA,SLPA)	Consensus clustering works well on soft clustering	Doesn't suits for hard clustering
A Survey of clustering ensemble algorithms[10]	Cluster-based similarity partitioning(CSPA), Hyper Graph partitioning Algorithm(HGPA), Meta-clustering Algorithm(MCLA)	To improve the performance quality.	Needs different types of algorithm to find the best solution

V. DISCUSSION

The necessary and importance of clustering is increasing day by day and in the paper is to explain how the clustering is working and how the consensus clustering is formed. It is just the review about cluster working in different areas and how it is useful when it is combined with other algorithm. The clustering algorithm has several advantages and disadvantages where some are mentioned below.

The advantages of k-means clustering is , it is easily understandable, fast to run, gives best result when datasets are distinct or well separated from each other. The disadvantages of k-means are the user's needs to specify the no of clusters and cluster points. It doesn't be able to solve when data's overlap each other. Each cluster will not have equal weight.

Hierarchical clustering has several advantages such as this algorithm doesn't need any aprior information about the nunber of clusters, it is easy to implement and gives best result except some cases. The main disadvantage of this algorithm is that it is difficult to handle for different sized clusters and sometimes it is difficult to identify the correct number of clusters.

The benefits of consensus clustering is that it produces better clustering, it finds combined clustering algorithm unattainable by any single clustering algorithm, less sensitive to noise, outliers or sample variations, it is able to integrate solutions from multiple distributed sources of data or attribute.

Here few real time medical datasets such as Breast Cancer, Ecoli, Iris, Dermatology, and Wine [1] are collected from the UCI repository. The execution time for the three algorithms are formed and plotted in the chart.

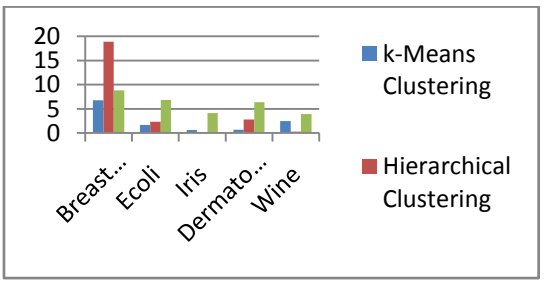

Figure 4: Comparison of 3 Clustering Algorithms

In figure 4 the execution time of clustering for different data set are shown. The Breast Cancer dataset has 699 object, Ecoli datasets has 332 objects, iris datasets has 150 objects, wine dataset has 178 object,

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

dermatology datasets has 358 objects. The above bar chart represents that iris and wine takes less time in hierarchical clustering where as Breast Cancer, Ecoli, Dermatology takes less time in k-means clustering. This proves that for small dataset hierarchical clustering performs well and for large dataset k- means clustering works well.

VI. CONCLUSION

As technology develops every day the data's that are stored are also increasing. The research for maintaining the data is the ongoing process and many new methods are developed to maintain the data in some order. The main aim of maintain the data is to reduce the storage area and to retrieve the data easily. Here we focus on the methods of working about k-mean, hierarchical and graph partition and the consensus clustering which produces optimized solution. The proposed work can be made as comparing all the clustering algorithms and find the best algorithm and it can be combined with consensus clustering to get the best result.

REFERENCES

[1].Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining Concepts and Techniques", 3rd ed. Elsevier.
[2].Junjie Wu , " Advances in k-means Clustering A Data Mining Thinking", Springer Theses Recognizing Outstanding Ph.D. Research.
[3].Partha Sarathi Bishnu and Vandana Bhattacherjee. "Software Fault Prediction Using Quad Tree Based K-Means Clustering Algorithm", IEEE Trans. On Know. And Data Engg., Vol.24, NO.6, June 2012.
[4].Xiaojun Chen, Xiaofei Xu, Joshua Zhexue Huang, and Yunming Ye, " TW-k-Means: Automated Two-Level Variable Weighting Clustering Algorithm for Multiview Data". IEEE Trans. On Know. And Data Engg., Vol.25, NO.4, June 2013.
[5].Xiao-Tong Yuan, Bao-Gang Hu, Senior Member, IEEE, and Ran He," Agglomerative Mean-Shift Clustering", IEEE Trans. On Know. And Data Engg., Vol.24, NO.2, June 2012.
[6].Junming Shao, Xiao He, Christian Bohm, Qinli Yang, and Claudia Plant," Synchronization-Inspired Partitioning and Hierarchical Clustering", IEEE Trans. On Know. And Data Engg., Vol.25, NO.4, June 2013.
[7].Guangxia Li, Kuiyu Chang, and Steven C.H. Hoi," Multiview Semi-Supervised Learning with Consensus," IEEE Trans. On Know. And Data Engg., Vol.24, NO.11, June 2012.
[8].Kunal Punera and Joydeep Ghosh," Consensus Based Ensembles of Soft Clusterings".IEEE Trans. On Know. And Data Engg., Vol.24, NO.11, June 2012.
[9].Jingxuan Li, Bo Shao, Tao Li, and Mitsunori Ogihara," Hierarchical Co-Clustering: A New Way to Organize the Music Data", IEEE Trans. On Know. And Data Engg., Vol.14, NO.2, April 2012.
[10].S.Vega-Pons and J. Ruiz-Shulcloper," A survey of clustering ensemble algorithms",Int.J.Pattern Recogn. Artif. Intell., Vol. 25, no. 3,2011.
[11]. Adriana Prado, Marc Plantevit, Ce line Robardet, and Jean-Francois Boulicaut , "Mining Graph Topological Patterns: Finding Covariations among Vertex Descriptors", IEEE Trans. On Know. And Data Engg., Vol.25, NO.9, September 2013.
[12]. Jianguo Wang, Eric Lo, and Man Lung Yiu,"Identifying the Most Connected Vertices in Hidden Bipartite Graphs Using Group Testing" , IEEE Trans. On Know. And Data Engg., Vol.25, NO.10, October 2013.