International Journal of Advanced Technology in Engineering and Science
Vol. No.4, Special Issue No. 01, February 2016
www.ijates.com
ijates
ISSN 2348 - 7550

K-MEANS BASED CONSENSUS CLUSTERING (KCC) A FRAMEWORK FOR DATASETS

B Kalai Selvi

PG Scholar, Department of CSE, Adhiyamaan College of Engineering, Hosur, Tamil Nadu, (India)

ABSTRACT

Data mining is the process of extracting the useful information from the clumsy of data. There are many portions in the data mining; one of the challenging part is Clustering. Clustering is the process of grouping the valid information from the raw data. The aim of this paper is to find the combination of K-Means algorithm and Consensus Clustering. the result of the KCC Algorithm is compared with K-means algorithm. We summarize the experiment and identify the common pitfalls in it.

Keywords: Data Mining, Clustering, K- means, Consensus Clustering

I. INTRODUCTION

Internet Contains large stuff of data which are not in the proper form. The data's are considered as the raw data. To make the raw data useful the data mining is used to extract the information in it[1]. Data mining contains multiple parts such as extracting, clustering, classification, and summarization. In this paper clustering is focused. Clustering is defined as the process of grouping the raw data in certain condition.

Clustering is classified as hard and soft clustering, sometime also classified as the exclusive clustering, overlapping clustering, Hierarchical Clustering, graph based clustering and so on[6] . In this paper we focus on the combination of two clustering algorithm to form the best performance according to time. The algorithms are K-Means algorithm and Consensus Clustering.

K-means Algorithm comes under soft clustering. Here more than one data can be grouped in two or more cluster. Consensus clustering is used to find the optimal solution from the multiple clustering[8]. To form the multiple clustering k-means is used. Thus the combination of these two algorithm is defined as K-means based Consensus Clustering(KCC).

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 architecture diagram and Section V discusses about the experiment and result. Finally, Section VI proposes some suggestions and conclusion.

II. RELATED WORK

Bishnu and Bhattacherjee [2] 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. [3] 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[9]. 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.

To improve the quality of individual data the methodology such as Cluster-based similarity partitioning (CSPA), Hypergraph partitioning Algorithm (HGPA), and Meta- clustering Algorithm (MCLA) is used[5]. 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 [4].

III. METHODOLOGY

3.1. K-Means Clustering Algorithm

The input data is clustered according to K-means algorithm[1]. The input data and the center of clusters are given as the input of the process. The data are clustered based on the distance metrics. Based on the small distance from the data to center the clusters are formed

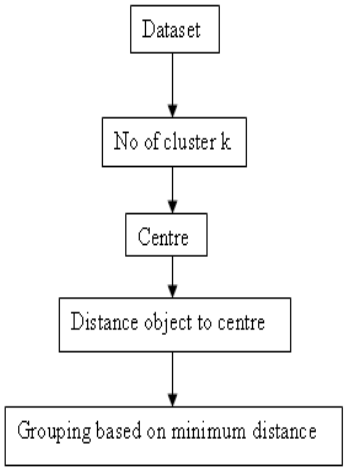

Fig 1 K-Means Algorithm

3.2. Consensus Clustering

The consensus clustering is used to find the optimal cluster. Any clustering algorithm can be used to form the optimal solution[7]. The clusters are stored in the contingency matrix.From the contingency matrix the best clustering is found

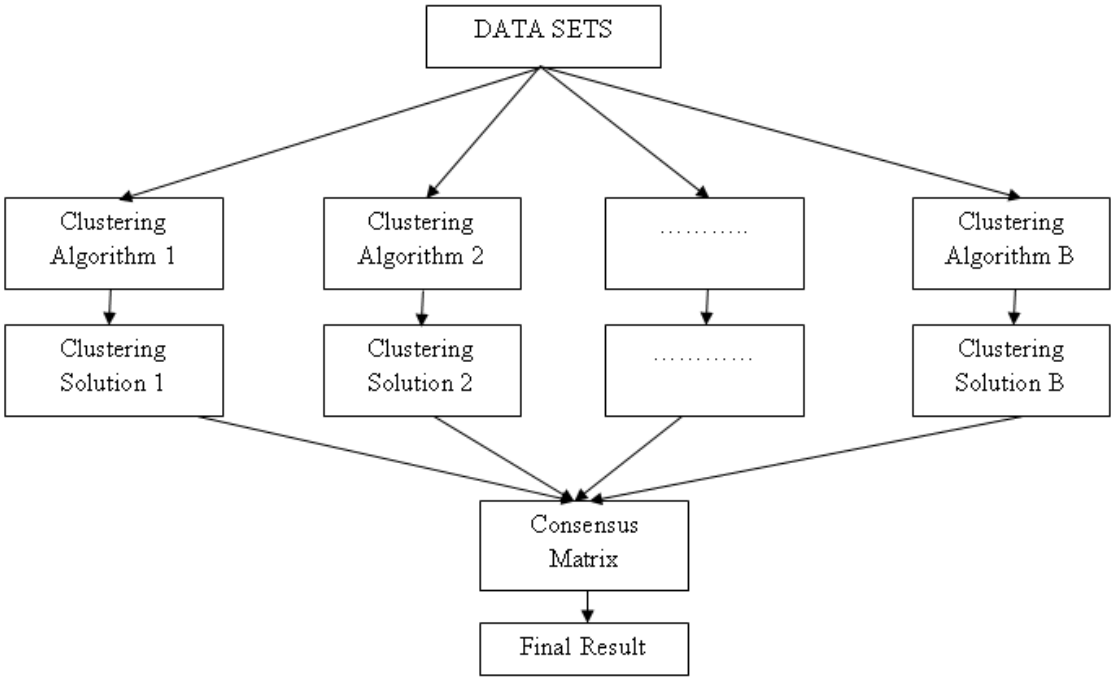

Fig 2 Consensus Clustering Algorithm

IV. ARCHITECTURE

The basic architecture that is followed in the project is given below:

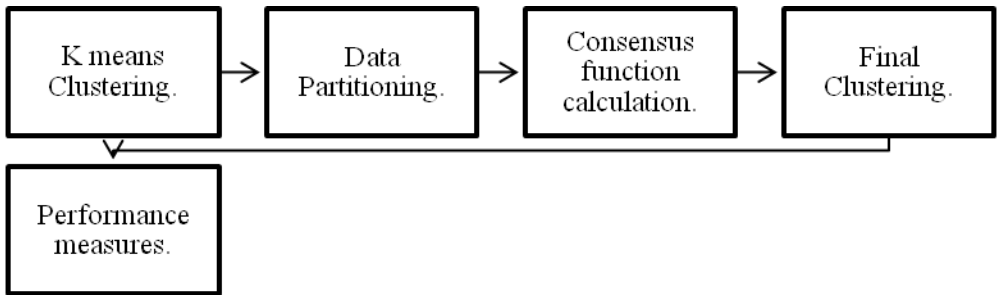

Fig 3 Architecture Diagram

V. EXPERIMENT AND RESULT

5.1. Data Partitioning

Before partitioning the missing data are removed from the dataset. The Partition of the data is done according to the data centers. The centers are given by the user to form cluster. The cluster is formed by using K-means algorithm.

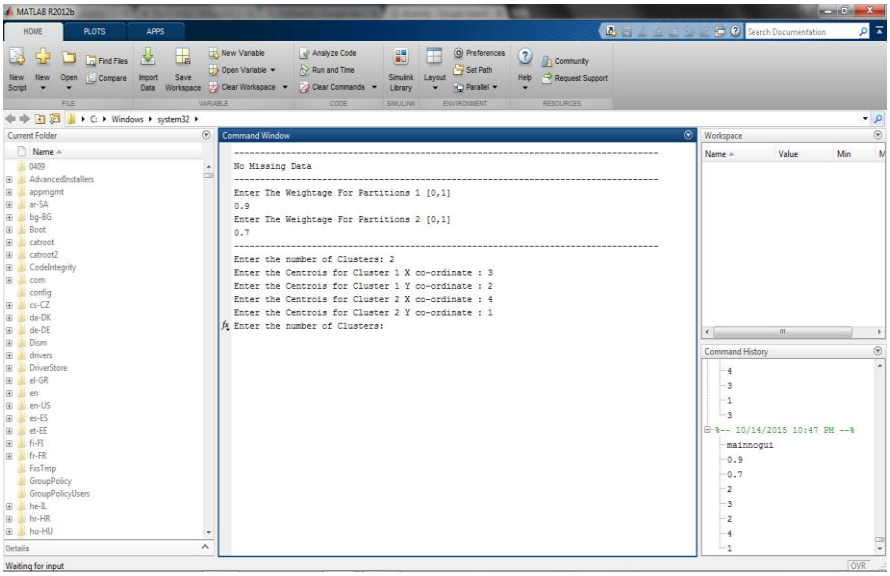

Fig 4 Data Partitioning input

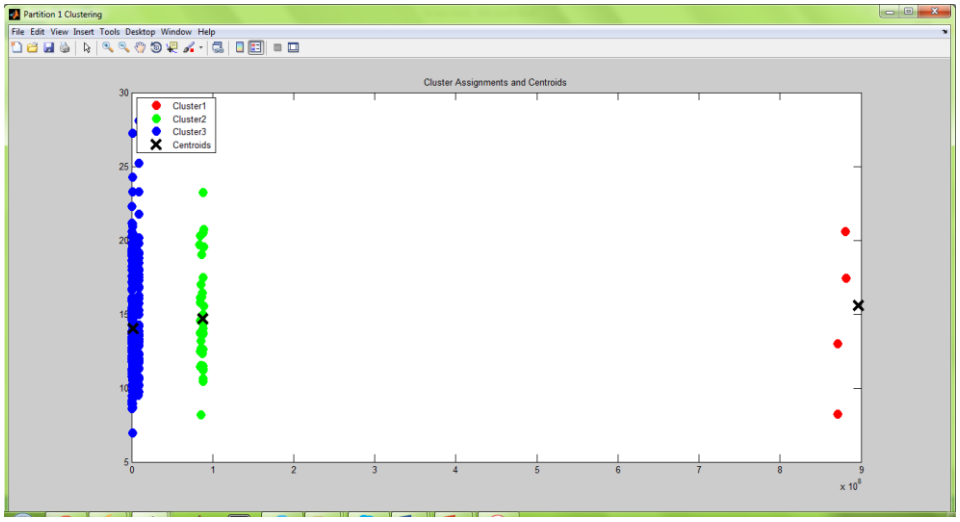

Fig 5 Data Partitioning Clusters

5.2. Consensus Clustering Function Calculation

To find the different metrics such as performance and improvement of quality the utility functions are used. The utility function is the combination of distance function with the input.

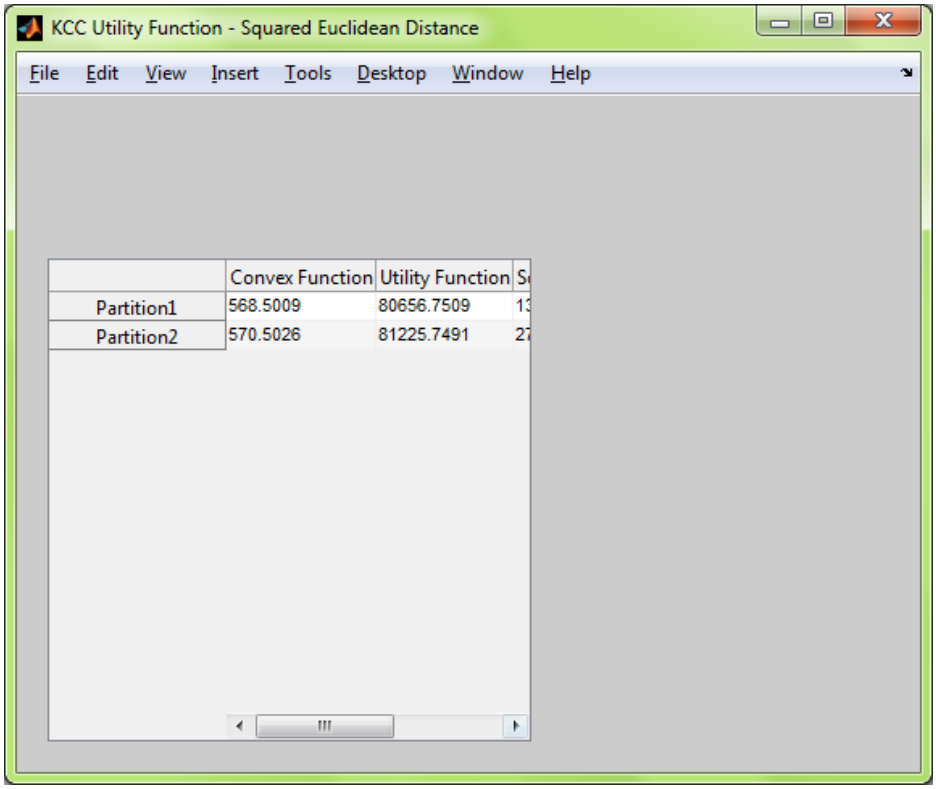

Fig 6 Consensus Clustering Function Calculation

5.3. Final Clustering Using Kcc

The utility function and other function are used to find the consensus clustering.The KCC is formed as the final clustering and results are given.

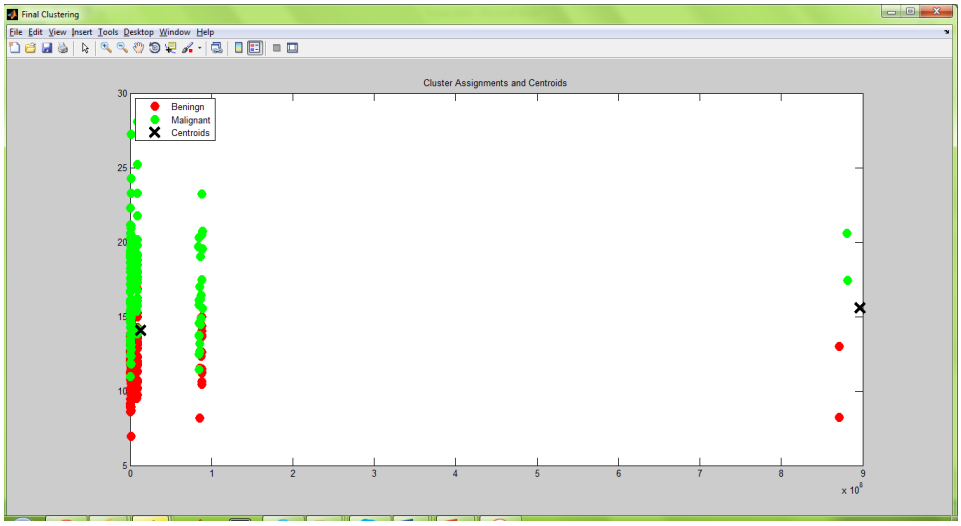

Fig 7 Final Clustering(KCC)

5.4. Performance Measure

The performance of KCC is compared with K-means algorithm. This helps to show that the performance of KCC is high than K-means in the execution time.

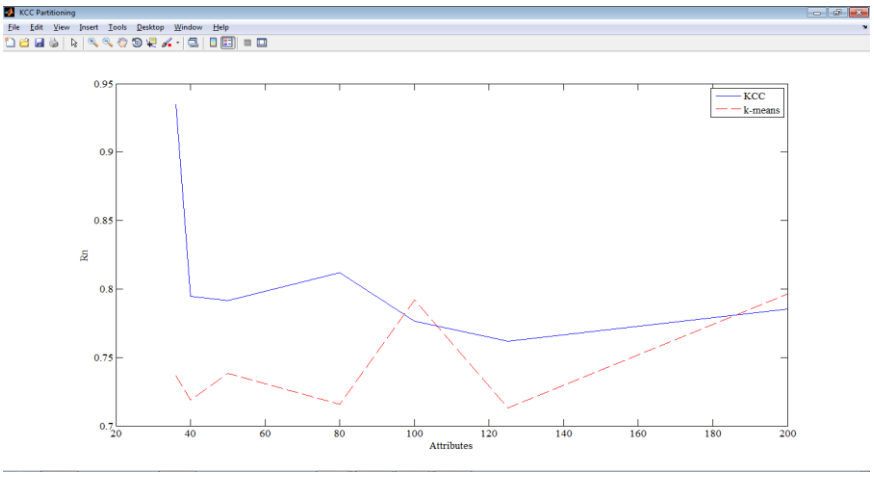

Fig 8 Comparision of KCC and K-means

VI. CONCLUSION AND FURTHER WORK

In this project the performance of K-means is compared with the K-means based Consensus Clustering based on execution time.It produces the result that KCC works faster than K-Means Algorithm.In future fuzzy c-means clustering is used to obtain the constant performance level when compared with k-means

REFERENCE

[1]. Junjie Wu,Hongfu Liu, Jie Cao, Jain Chen," K-Means-Based Consensus Clustering: A Unified View",IEEE Trans. Knowl. Data Engi.,vol. 27,no. 1,pp 155-169.

[2]. Partha Sarathi Bishnu and Vandana Bhattacherjee," TW-k-Means: Automated Two-Level Variable Weighting Clustering Algorithm for Multiview Data", IEEE Trans. On Know. and Data Engg., vol. 24,no.6, April 2013.

[3]. Xiaojun Chen, Xiaofei Xu, Joshua Zhexue Huang, and Yunming Ye," Software Fault Prediction Using Quad Tree-Based K-Means (QDK) Clustering Algorithm", IEEE Trans. On Know. And Data Engg., vol24,no.6, June 2012

[4]. S.Vega-Pons and J. Ruiz-Shulcloper," A survey of clustering ensemble algorithms",Int.J.Pattern Recogn. Artif. Intell., Vol.25, no. 3,2011.

[5]. Xiao-Tong Yuan, Bao-Gang Hu and Ran He," Agglomerative Mean-Shift Clustering", IEEE Trans. On Know. And Data Engg., vol24,no.2, Feb. 2012

[6]. Sandro Vega-Pons and Jose Ruiz-Shulcloper," A Survey of clustering ensemble algorithms", International Journal of Pattern Recognition and Artificial Intelligence Vol. 25, No. 3 (2011)

[7]. Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining Concepts and Techniques", 3rd ed. Elsevier

[8]. Junjie Wu , " Advances in k-means Clustering A Data Mining Thinking", Springer Theses Recognizing Outstanding Ph.D. Research

[9]. 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.