International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 07 | July -2017 www.irjet.net p-ISSN: 2395-0072

Two Layer k-means based Consensus Clustering for Rural Health Information System

Ms. Archana Singh[1], Prof.Dr.V.H.Patil[2]

[1]PG scholar, Department of computer Engineering, Matoshri College of Engineering and Research Centre, Maharashtra, India
[2]Professor, Department of computer Engineering, Matoshri College of Engineering and Research Centre, Maharashtra, India

---

Abstract - *This paper presents a data clustering approach using two layer k means based consensus clustering. This algorithm helps to partitions the heterogeneous data and form the sub clusters for main clusters to find efficient decision making in rural health information system. In this paper we provide a systematic study of two layer k means based consensus clustering. It helps in both complete and incomplete datasets. Experimental results on MCTS dataset demonstrate that two layer k-means based consensus clustering is highly efficient and comparable to the rural health information system. This algorithm shows high robustness to incomplete basic partitioning with many missing values.*

Key Words: Consensus clustering, k-mean. Sub cluster, classification

1. INTRODUCTION

In Traditional K-means algorithm, a datum X will be assigned to the cluster C where the distance between X and the cluster center of C is minimal, comparing to the distances between X and the cluster centers of other clusters. However, the abnormal data may be assigned to most of clusters but normal data are classified into a few clusters.

Consensus clustering, also known as cluster ensemble or clustering aggregation, aims to find a single partitioning of data from multiple existing basic partitioning [3]. It has been widely recognized that consensus clustering can help to generate robust clustering results, handle noise, outliers and sample variations, and integrate solutions from multiple distributed sources of data or attributes [4].

The main objective to develop this system that helps, the rural officers to store and upload the data directly to the cloud server which will help them in reducing the paper work. According to the existing scenario the rural officers go and collect the information of the rural infants and pregnant lady to provide them with proper vaccination. The details are collected and after few days the details are stored into the server and the further process of vaccination is carried out, this process take time to get the output. The vaccination is provided to the pregnant lady till the birth of baby with a regular period of time. The required medicine is also provided to the babies till the age of 5 years.

This project is provides proper messages to the families about the necessary of taking the vaccination at proper time, this application will provide the notifications to the families about the date, time and location of the vaccination facility provided by the government. This will help in getting the proper count of infants and pregnant ladies in a village those who have taken the vaccination which will help the rural officer to know who has taken and number of counts remaining to take the vaccination from a particular village.

It will notify the families about the funds that are released from the government to the pregnant ladies and to new born babies in rural area for the pregnancy and for further medical support. This help in improving the nutritional and health status of children's in the age group 0-5 years to enhance the capability of the mother to look after the normal health and nutritional needs of the child through the proper nutrition and through health education.

To develop this system we are going to use two layer k-means based Consensus clustering to take decision on the scattered or heterogeneous data. Before work on this algorithm first we know about basic concepts of Consensus Clustering and two layer K-means.

2. REVIEW OF LITERATURE

This chapter presents the study of the existing systems and proposed work related to the proposed System. The purpose of the literature survey is to identify information relevant to project work and the potential and known impact of it within the project area.

Rural Health Information System is an existing system which works on a mission of providing the information, to the rural areas about the vaccination required for infants and pregnant women. The Rural Health Information System has rather had adverse impact on health system, all services are paralyzed and health supervisors are sitting in front of office computers and making data entries. Currently Rural Health Information System portal is an absolute online version and for poor internet connectivity it cannot be operated properly. Adding a feature for offline data entry process followed by online uploading to the server would be useful. As no IEC activity can be conducted during data entry

period, which may lead to some primary health issues. Till the data entry work outsourcing is done, they are promoting record maintenance at overall primary health care services. Regular feedback system will be established in this system. This system will be in a position to benefit the entire health care system incorporating more useful indicators. Hence, we summarized that the existing system is inefficient, time consuming, poorly managed, and lacking flexibility. This solutions very useful as the solution is inherently distributive.

3. SYSTEM OVERVIEW

3.1 Consensus Clustering

Consensus clustering, also known as cluster ensemble or clustering aggregation, aims to find a single partitioning of data from multiple existing basic partitioning [3]. It has been widely recognized that consensus clustering can help to generate robust clustering results, handle noise, outliers and sample variations, and integrate solutions from multiple distributed sources of data or attributes [4]. Here we briefly introduce the basic concepts of consensus clustering and formulate the research problem of this paper.

Let X = {x1, x2, ⋯ ,xn} denote a set of data objects/points/instances. A partitioning of X into K crisp clusters is represented as a collection of K subsets of objects in C = {Ck|k = 1,⋯ ,K}, with CkTCk0 = ∅, ∀k 6= k0, and SK k=1 Ck = X, or as a label vectorπ = hLπ(x1),⋯ ,Lπ(xn)i, where Lπ(xi) maps xi to one of the K labels in {1,2,⋯ ,K}. We also use some conventional math notations as follows. For instance, R, R+, R++, Rd, and Rnd denote the sets of reals, non-negative reals, positive reals, d-dimensional real vectors, and n×d real matrix, respectively. Z denotes the set of integers, and Z+, Z++, Zd and Znd are defined analogously. For a d-dimensional real vector x, kxkp denotes the Lp norm of x, i.e., kxkp = p qPd i=1 xp i , |x| denotes the cardinality of x, i.e., |x| =Pd i=1 xi, and xT denotes the transposition of x. The gradient of a single variable function f is denoted as ∇f, and the logarithm of based 2 is denoted as log. In general, the existing consensus clustering methods can be categorized into two classes, i.e., the methods with or without global objective functions [9]. In this paper, we are concerned with the former methods, which are typically formulated as a combinatorial optimization problem as follows. Given r basic partitioning of X (a basic partitioning is a crisp partitioning of X by some clustering algorithm) in Π = {π1,π2,⋯ ,πr}, the goal is to find a consensus partitioning π such that

$$\Gamma(\pi, \Pi) = \sum_{i=1}^{r} w_i U(\pi, \pi_i) \quad (1)$$

is maximized, where Γ : Zn++ × Znr++ 7→ R is a consensus function, U : Zn++×Zn++ 7→ R is a utility function, and wi ∈ [0,1] is a user-specified weight for πi, with Pr i=1 wi = 1. Sometimes a distance function, e.g., the well-known Mirkin distance [5], rather than a utility function is used in the consensus function. In that case, we can simply turn the maximization problem into a minimization problem without changing the nature of the problem. Consensus clustering as a combinatorial optimization problem is often solved by some heuristics and/or some Meta heuristics. Therefore, the choice of the utility function in Eq. (1) is crucial for the success of a consensus clustering, since it largely determines the heuristics to employ. In the literature, some external measures originally proposed for cluster validity have been adopted as utility functions for consensus clustering, e.g., the Normalized Mutual Information [3], Quadratic Mutual Information [6], and Rand Index [8]. These utility functions of different math properties pose computational challenges to consensus clustering.

3.2 Two Layer K-mean Algorithm

In the traditional K-means algorithm, a datum X will be assigned to the cluster C where the distance between X and the cluster center of C is minimal, comparing to the distances between X and the cluster centers of other clusters. However, if there are outliers or noisy data in a data set, the abnormal data may be assigned to most of clusters but normal data are classified into a few clusters. In Fig. 1, the red dots are the abnormal data which are only a few data relative to normal data and are classified to two clusters but the vast majority of normal data are classified to only one cluster. It is usually non-helpful for future analysis. In data clustering, when the data in one cluster are quite different, the respective features of the cluster cannot precisely represent all the data in this cluster. According to this requirement, a two-layer K-means algorithm is proposed to improve traditional K-means algorithm. The two-layer K-means algorithm contains three steps:

A. Data normalization,
B. Cluster center initialization, and
C. Two-layer clustering.

A. Data normalization

In distance-based classification, a small variation in one feature is probably more influencing than a big variation in other feature when computing the distance of two data. It is necessary to normalize every feature value of each feature dimension to a specific range. This stage is to transform all variables in the data to a specific range. Let S={X1, X2, ..., XN} be a data set consisting of N data, Xi be the i-th data in S, and (xi1, xi2, ..., xid) be the features of Xi. For each feature value xid is normalized by (2).

$$x'_{ij} = \frac{x_{ij} - \underset{k=1}{\overset{N}{MIN}}(x_{kj})}{\underset{k=1}{\overset{N}{MAX}}(x_{kj}) - \underset{k=1}{\overset{N}{MIN}}(x_{kj})} \quad (2)$$

B. Initial cluster center

In this step, a most discrepant initial cluster center method is proposed to determine the initial cluster centers for K-means algorithm. It uses the biggest discrepant data as the initial cluster centers. Let the distance dik of two data C1 and C2:

$$d_{ik} = \sum_{j=1}^{d} w_j \mid x'_{ij} - x'_{ki} \mid^{r_j} \quad (3)$$

where wj and rj are the given constants. The algorithm first decides two data C1 and C2, where C1=Xi and C2 = Xk, and

$$(i,k) = \arg\left(\max_{i=1}^{N-1} \left(\max_{k=i+1}^{N} d_{ik} \right) \right) \quad (4)$$

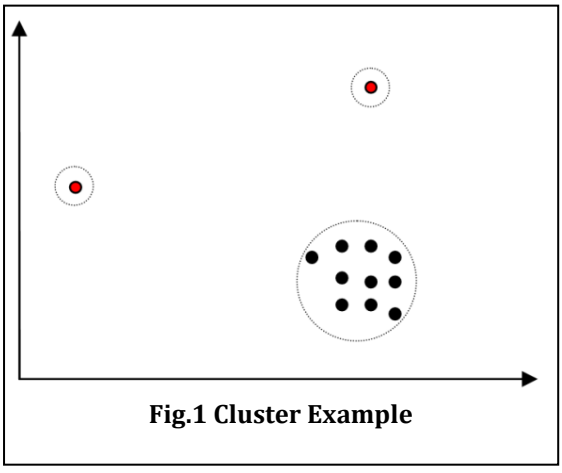


Fig.1 Cluster Example

After that, it computes the data: C3 which is farthest from C1 and C2, C4 which is farthest from C1, C2 and C3, CK which is farthest from C1, C2, …, and CK-1, where C1, C2, …, and CK are in S and are considered to the initial cluster centers of the K clusters.

C. Two-Layer Clustering

K-means algorithm uses a cluster center to represent the data of the cluster. If the dissimilarity of data is big in a cluster, the cluster center cannot describe all of the data in the cluster. For example, there are two clusters C1 where the dissimilarity of data is very big and C2 where the dissimilarity of data is very small. Then enter a data point x which is closer to the border of C1 but far from C2. On the contrary, x is far from the cluster center of C1 but closer to C2.

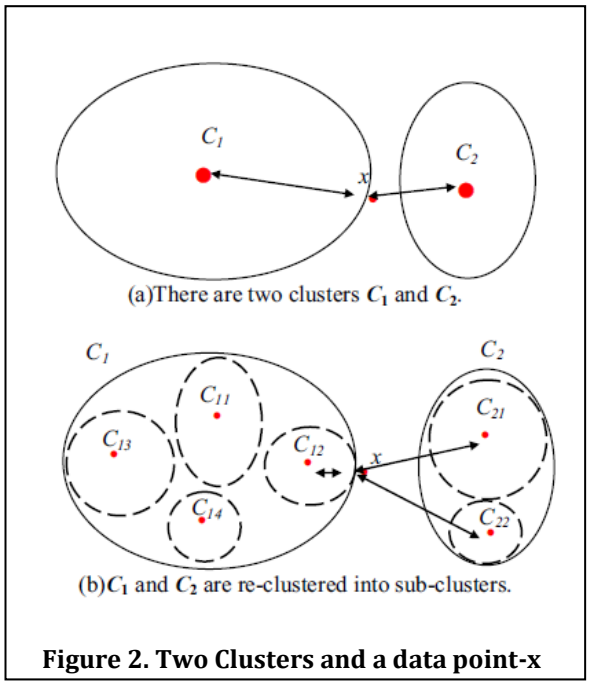


Figure 2. Two Clusters and a data point-x

As Fig. 2(a), x will be mistakenly classified to the cluster C2 in the traditional K-means algorithm. However, when the data in C1 are separated into several smaller clusters (sub clusters), x will be pointed to one of sub-clusters in the cluster C1 as Fig. 2(b) so x will definitely belong to cluster C1. This study proposes two-layer K-means algorithm to improve the traditional K-means algorithm due to the above problems. Each cluster is subdivided into several sub-clusters in two-layer K-means algorithm and then combines with the traditional K-means algorithm for data clustering.

Firstly, two-layer K-means algorithm adopts the data normalization step and the initial cluster center step and then cluster data set S into K clusters using traditional K-means algorithm. Next assuming that CG is the G-th cluster of data set S, and two-layer K-means algorithm uses traditional K-means algorithm to divided the CG into KG sub-clusters with CG1, CG2, …, G GKC . To input a data point (v1, v2… vd), the data point is detected belongs among certain sub-clusters of CG , then it will be attributed to an element of CG. Next let CGgj is the j-th dimension value of g-th sub-cluster center in the G-th cluster, the distance dis between the data point and each sub-cluster center as follows:

$$dis = \underset{g=1}{\overset{K_G}{MIN}} \sum_{j=1}^{d} w_j \mid C_{Ggj} - v_j \mid^{r_j} \quad (5)$$

The sub-cluster is not only closest to the data point, but also belongs to the cluster CG, and then the data point is classified into CG.

4. SYSTEM ANALYSIS

In proposed system we will consider two one is two layer k-means and second is consensus clustering. It will help to make decision on heterogeneous data which were not considered in existing system. In this approach heterogeneous data is considered as input and are used to test and validate the results.

As Shown in figure 3, the heterogeneous data is taken as a input to system. This data is collected through Asha's daily collection of information from different regions. After that Consensus clustering is applied on this data to make basic clusters. On these clusters two layer k-mean algorithm is applied for making clusters of sub clusters. These are nothing but the more nearest to neighbor solution which are used to gain the efficient results.

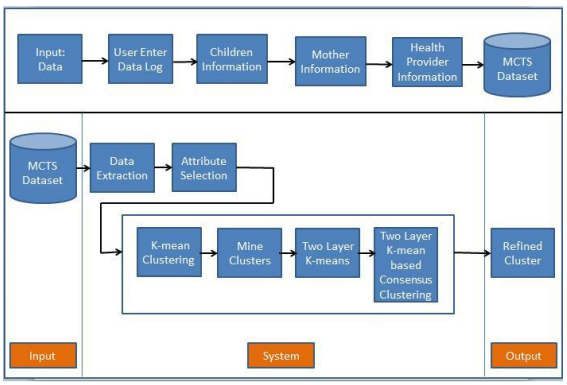


Figure 3. Proposed System Architecture

Fig. 4 explains that the flow of the MCTS using two layer k-means based Consensus clustering to make effective and correct decision on heterogeneous data to send notifications or textual messages to pregnant women for their better health.

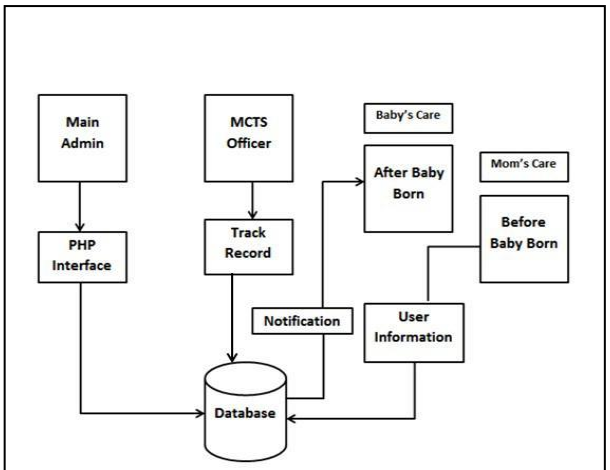


Figure 4. Functional model of proposed System

In this our system there are three basic parts or module:

1) Main admin
2) MCTS Officer
3) User Side (Baby's care and Mom's Care

In first part i.e. Main admin has web panel to handle the following tasks, admin has rights to allocate MCTS officer in particular part of country and track the work of MCTS officers and the data inserted by the officers. The admin has control to send proper notification to user.

In second module, the MCTS officers also called ASHA have to work on field. They are having application to register pregnant women or new born baby. This will reduces the paper work of MCTS officer which they are doing currently.

Third module is important i.e. User Application which will be used by end users. These users include pregnant women and new born baby mothers. Pregnant women will get proper notification for dosage and there timing during pregnancy period. New baby born mothers will get notifications for vaccinations.

In this system we are going to apply the two layer k-means based consensus clustering algorithm on the data gathered by MCTS officers. The notification will fired on the decision of algorithm applied on that data.

5. MATHEMATICAL MODELLING

Let S be a technique for two layer k means based consensus clustering; the equation proposed will be from the fundamental principles of k means and consensus clustering.

Input:

Cj represents the set of clusters

CKj represents the set of Sub Clusters
Where,

Cj = c1,c2,c3...cn

CKj = ck1,ck2,ck3...ckn

Output:

: The consensus partitioning

Process:

S={ D, X(b), K, pi}

Where,

S= System.

D= Healthcare Dataset

Where Cj, CKj are arbitrary dataset which is actual input for system
X (b) = clusters

K=Two layer k means based consensus Approach
Km = k-means clustering
2Km= 2 layer k means

pi= The consensus partitioning

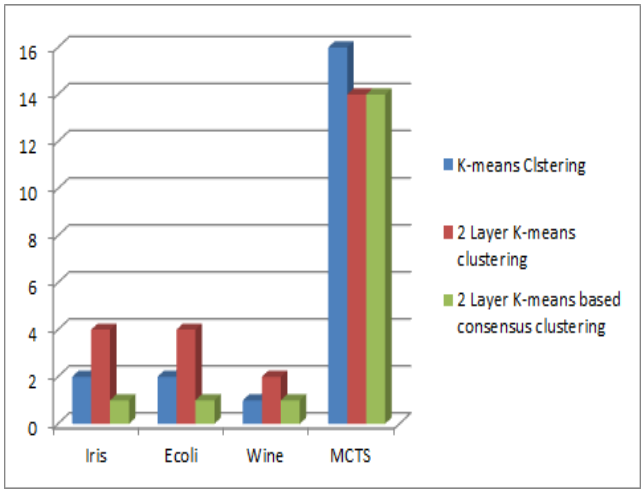
6. RESULT ANALYSIS

After comparing the classifier performance against all three data mining model, quiet interesting results were discovered as shown below. We used a test bed consisting of a number of real world data sets obtained from UCI repositories like Iris, Ecoli and Wine.

Table 1 shows the complete summary of the performance comparisons as execution time of the three data mining models used for this research work on different datasets. Also Table 2 shows the complete summary of the performance comparisons as accuracy of the three data mining models used for this research work on different datasets.

Table 1: Comparison of Execution Time (in ms)

Algorithm	Iris	Ecoli	Wine	MCTS
K means	2	2	1	16
2 layer k means	4	4	2	14
2layer k-mean Consensus	1	1	1	14

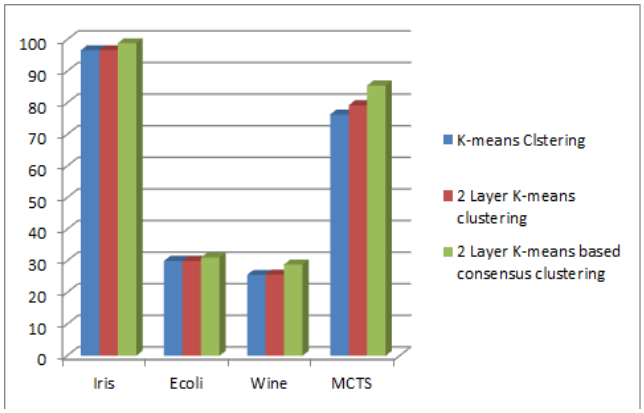


We do the comparison of Two layer K-mean consensus clustering algorithm with K-mean and Two layer k-mean in terms of execution efficiency. Table 1 shows the runtime comparison of the three methods where we observe that proposed algorithm two layer k-mean based consensus clustering algorithm is fastest among three methods.

Also table 2 demonstrated that accuracy of proposed algorithm is higher than the other algorithms in different datasets.

Table 2: Comparison of Accuracy (in percentage)

Algorithm	Iris	Ecoli	Wine	MCTS
K means	96.67	30.01	25.6	76.25
2 layer k means	96.68	30.02	25.7	79.24
2layer k-mean Consensus	98.9	31.06	28.9	85.51



We established the general theoretical framework of two layer k-means based consensus clustering and provided the corresponding algorithm. Experiments on real world datasets have demonstrated that two layer k-mean based consensus clustering has high efficiency and shows the robust performances.

7. CONCLUSION

In this paper, two layer k-means consensus algorithm is proposed to improve the two layer k-means and traditional k-means algorithm. It can give a better accuracy rate of data clustering than other k-means algorithm. The two layer k-means consensus algorithm is used to strengthen the accuracy of data clustering.

Two layer k-means based Consensus Clustering for Rural health information system used for classifying the groups and sending the vaccination notification in the form of textual messages and reminders to the families of infants and pregnant women, according to the vaccination dates periodically by using their registered identification number.

8. ACKNOWLEDGEMENT

I want to thank all people who help me in different way. Especially I am thankful to my guide and HOD Prof.Dr.V.H.Patil for her continuous support and guidance in my work.

REFERENCES

[1] Junjie Wu, Member, IEEE, Hongfu Liu, Hui Xiong, Senior Member, IEEE, Jie Cao, Jian Chen, Fellow, IEEE,"K-means based Consensus Clustering: A Unified View",IEEE Transaction On Knowledge And Data Engineering, vol. xx, no. xx, December 2013.

[2] Chen-Chung Liu1,Shao-Wei Chu2, Shyr-Shen Yu4,*,Yung-Kuan Chan3,"A modified K-means Algorithm- Two Layer K-means Algorithm",2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing.

[3] A. Strehl and J. Ghosh, "Cluster ensembles — a knowledge reuse framework for combining partitions," JMLR, vol. 3, pp. 583–617, 2002.

[4] N. Nguyen and R. Caruana, "Consensus clusterings," in ICDM, 2007.

[5] B. Mirkin, "The problems of approximation in spaces of relationship and qualitative data analysis," Information and Remote Control, vol. 35, p. 1424–1431, 1974

[6] A. Topchy, A. Jain, and W. Punch, "Combining multiple weak clusterings," in ICDM, 2003, pp. 331–338.

[7] Z. Lu, Y. Peng, and J. Xiao, "From comparing clusterings to combining clusterings," in AAAI, 2008, pp. 361–370.

[8] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for k-means clustering," in KDD, Paris, France, 2009, pp. 877–886.

[9] J. Wu, H. Xiong, C. Liu, and J. Chen, "A generalization of distance functions for fuzzy c-means clustering with centroids of arithmetic means," TFS, vol. 20, no. 3, pp. 557–571, 2012.

[10] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Addison-Wesley, 2005.