ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Fuzzy C-mean Clustering Using Randomized Dimensionality Reduction

Shilpa Mokashe, Prof. Sanjay B. Thakare,

M. E Student, Dept. of Computer Engineering, RSCOE, Pune, Maharashtra, India

Assistant Professor, Dept. of Computer Engineering, RSCOE, Pune, Maharashtra, India

ABSTRACT: In dimensionality reduction method, feature extraction and feature selection are the two strategies are available. A small subset is choose from provided features for feature selection that is based on chi square algorithm and implements k-means clustering over the features. System implements SVD random projection algorithm for extraction of feature. Previous system implemented k-means clustering algorithm for dimensionality reduction but this algorithm have few disadvantages in the system. Algorithm is slower in speed as well as accuracy also less. So, to avoid this in proposed system, we utilized fuzzy c-means algorithm for the clustering procedure. C-means algorithm gives accurate outcomes as compared with the k-means clustering algorithm.

KEYWORDS: Dimensionality reduction, fuzzy C-means clustering algorithm, clustering, k-means clustering.

I. INTRODUCTION

The process of dividing data from unique in group is called as data clustering. Thus, contents inside the group that as similar as possible and contents in different groups are as distinct as may be expected normally. Clustering is implemented on the basis of the manner of the data and number of computations of similarity is implemented to place content in the groups, where the similarity computation controls over formation of the groups. Some examples of calculations are implemented as a part of clustering such as including separation, network and energy [7].

Data is divided in specific cluster in complex clustering, where each data element has accurate place in single cluster. Fuzzy clustering allows data elements place in several clusters and also associated with each element is a set of membership levels.

Number of applications includes clustering such as in science as well as in engineering. Additionally, clustering implemented in bioinformatics as well as medication to the sociologies and internet. Most popular clustering algorithm is called as “C-means” algorithm or Lloyd’s method [1]. Lloyd's technique is an iterative desire-augmentation type methodology that attempts to confront the going with target. Amount of cluster partitions the point into k clusters. So the aggregate entire of the squared Euclidean distances of each point to its closest cluster focus is diminished. In view of this insightful objective and also its sufficiency, the Lloyd's framework for k-means clustering has ended up being colossally surely understood in applications. Given a set of Euclidean point and a positive entire number k-means starting late, the high dimensionality of present day tremendous datasets has given a critical test to the structure of powerful algorithmic reactions for k-means packing.

Furthermore we will see in this paper: Section II describes related work analysed topic. Section III presents current implementation details, introductory definitions and documentations and also properly illustrates the proposed system experiment tended to by this paper. Section IV demonstrates conclusions.

II. RELATED WORK

This section describes previous work accomplished by the researchers for text mining procedure. Christos Boutsidis, AnastasiosZouzias, Michael W. Mahoney, and PetrosDrineas [1] studied the concept of dimensionality reduction for k-means clustering. Dimensionality reduction incorporates the union of two methodologies: 1) feature selection and 2) feature extraction. A feature selection-based algorithm for k-means clustering selects a little subset of the input features and afterward applies k-means grouping on the selected features. An feature extraction-based algorithm for k-means clustering builds a little set of new counterfeit features and afterward applies k-implies bunching on the built features

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

[1]. Regardless of the significance of k-means clustering and the wealth of heuristic systems addressing to it, provably accurate feature determination systems for k-means clustering are not known. On the other hand, two provably accurate component extraction systems for k-means clustering are known in the literature; one depends on random projections and the other depends on the singular value decomposition (SVD) [10].

Christos Boutsidis, Michael W. Mahoney, and PetrosDrineas [2] present a feature selection algorithm for the k-means clustering issue. This algorithm is randomized and, accepting an exactness parameter $\epsilon \in (0, 1)$, selects and properly rescales in an unsupervised way $\Theta(k \log(k/\epsilon)/\epsilon 2)$ highlights from a dataset of arbitrary measurements. They demonstrate that, if they chance that they run any γ-approximate k-means calculation ($\gamma \geq 1$) on the features chose utilizing their technique, they can discover a $(1 + (1 + \epsilon)\gamma)$- approximate partition with high probability.

S. Arora, E. Hazan, and S. Kale [3] depict a direct sporadic analysis based technique for making lacking network approximations. Their system and analysis are to a excellent degree essential: the analysis uses basically the Chernoff - Hoeffding limits. Notwithstanding the straightforwardness, the theory is practically identical likewise, in few time superior to anything previous work. Their estimation enrolls the inadequate structure estimation in a singular overlook the data. Next, extensive zones in the output framework are quantize additionally might be succinctly related to a bit vector, in this way provoking much save stores in space.

K. L. Clarkson [4] present the Johnson-Lindenstrauss sporadic projection lemma gives an essential way to deal with manage decrease the dimensionality of a course of action of concentrates while around ensuring their pair wise partition. The clear utilization of the lemma connected to an obliged course of action of concentrates, however recent work has added to the framework to relative subspaces, curves and regular smooth manifolds.

S. Lloyd [5] proposed that they perceive two problems fused into incorporating to an automated segment subset determination calculation for unlabelled information. The essential for finding the measure of social occasions in conjunction with highlight choice and the essential for finding the amount of highlight determination criteria as to estimation. They investigate the part choice problem and this problem by FSSEM (Feature Subset Selection utilizing Expectation-Minimization (EM) pressing) and by two specific execution criteria for studying confident segment subsets: encrypt distinctness and most convincing possibility.

III. PROPOSED SYSTEM

This section describes the proposed system architecture in detail, system overview, proposed algorithm, mathematical model of the proposed system.

Problem Statement:

In the existing system the high dimensionality of modern massive datasets has provided a considerable challenge to the design of efficient algorithmic solutions for k-means clustering is: Ultra-high dimensional data force existing algorithms for k-means clustering to be computationally inefficient. The existence of many irrelevant features may not allow the identification of the relevant underlying structure in the data.

A. *System Overview*

The figure 1 shows the architectural view of the proposed system. The description of the system is as follows :

In this system initially input dataset is uploaded. Dataset contain number of features extraction and feature selection is done. The process of fuzzy C-means clustering is performed and final result is obtained.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

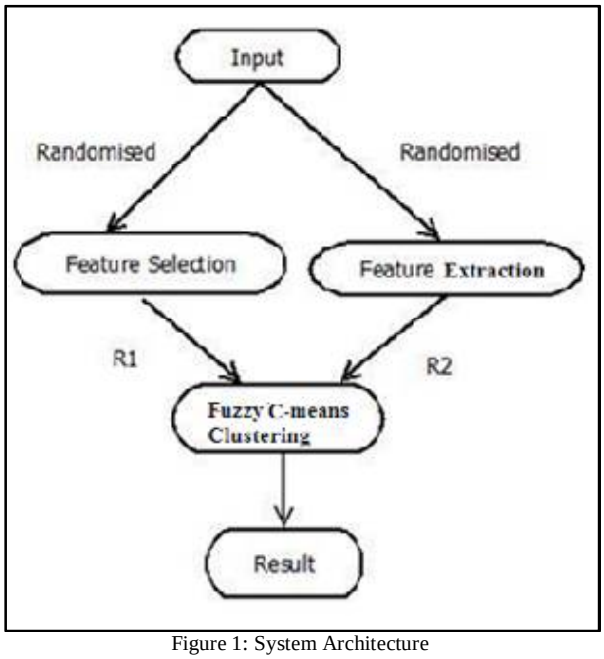

Figure 1: System Architecture

B. *Mathematical Model*

System S is represented as S= {I, F, FE, X, A}

Input

Input Processing

Dataset for input is Spellman dataset.

I= {i1, i2, i3, ...,in}

Where I is the set of Input (Gene Expression) and i1, i2, i3, ..., in are the number of inputs(Gene ids).

Process

1) Feature Selection

F = {f1, f2, f3, ...,fn}

Where F is the set of Feature Selection and f1, f2, f3, ...,fn represent as a number of selected features.

2) Feature Extraction

FE = {fe1, fe2, fe3, ...,fen}

Where FE is the set of Feature Extraction and fe1, fe2, fe3, ...,fen represent as a number of extracted features.

3) Fuzzy C-Mean Clustering

Let X = {x1, x2, x3 ..., xn} be the set of data points and V = {v1, v2, v3 ..., vc} is set of centroids .

a) In First iteration select 'c' cluster centersRandomly.

b) Calculate the fuzzy membership 'µij ' using:

$$u_{ij} = 1/\sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

c) find fuzzy centers 'V_j ' using:

$$V_j = \left(\sum_{i=1}^{n} (\mu_{ij})^m X_i\right) / \left(\sum_{i=1}^{n} (\mu_{ij})^m\right), \forall_j = 1,2,.....C$$

d) Repeat step b) and c) until the minimum 'J' value is achieved

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

where
'k'= iteration step.
'J'= objective function.

4) The SVD of a m-by-n matrix A is given by the formula:

$$A = UWV^T$$

Where
U means m-by-n matrix of the ortho normal eigenvectors of AA^T
V^T is the transpose of a n-by-n matrix containing the orthonormal eigenvectors of $A^T A$
W means n-by-n Diagonal matrix of the singular values which are the square roots of the eigenvalues $A^T A$

C. *Algorithms*
Algorithm 1 Proposed System Algorithm

1: Set the number of clusters, the fuzzy parameter (a constant > 1), and the stopping condition
2: Initialize the fuzzy partition matrix
3: Set the loop counter k = 0
4: Calculate the cluster centroids
5: For each data point, for each cluster, compute the membership values in the matrix
6: between consecutive iterations is less than the stopping condition, then stop; otherwise, set k=k+1 and go to step 4
7: Clustered data points

Algorithm 2 Randomized Feature Selection for k-Means Clustering

Input: Dataset $B \in Qj \times i$, no of clusters h and ε is in between 0 to 1/3.
Output: $D \in Qj \times r$ with s = O (h log(h)/ ε 2) attributes.
1: Let X be the SVD(B, h, ε); $X \in Qj \times h$
2: Let s = t1 4h ln(200h)/ ε 2
3: Let [M] = RandomizedSampling(X,s);
4: Return $D = BM \in Qj \times s$ with s rescaled columns from B

IV. RESULTS AND DISCUSSION

This section describes the result obtained by the proposed. The accuracy result and speed obtained by the fuzzy c-means clustering algorithm is more than the k-means clustering algorithm.

Figure 2 demonstrates the time comparison between fuzzy c-means clustering algorithm and k-means clustering algorithm. Fuzzy c-means algorithm consumes less time as compared to k-means algorithm and save the time.

Figure 3 graph demonstrates the accuracy comparison between the fuzzy c-means and k-means clustering algorithm. The accuracy outcome of the fuzzy c means clustering algorithm is more than the k-means clustering algorithm.

Advantages:

1) Proposed algorithm is best in case of accuracy as well as time.
2) In previous clustering algorithm, data point related to only one cluster point. Similar pointed data connection of every cluster center which can be related with one cluster centre.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

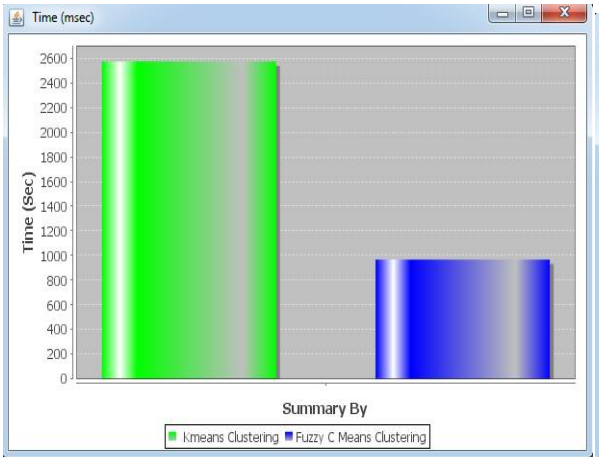

Figure 2: Time Comparison graph

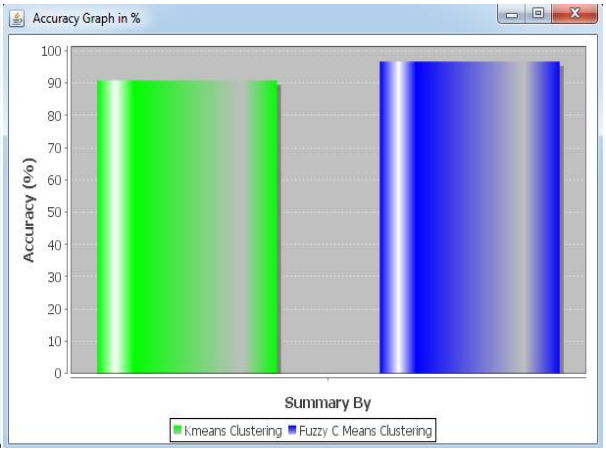

Figure 3: Accuracy Comparison graph

V. CONCLUSION AND FUTURE WORK

The proposed system focuses over the problem of dimensionality reduction of clustering. Number of existing systems in this subject includes heuristic approaches in the outcome that experimental performance will not be explained with a theoretical analysis. Dimensionality reduction technique works over excellent in principle. We present three approaches such as selection of feature technique for clustering as well as two techniques for extraction of feature techniques. Additionally, system utilizes fuzzy C-means clustering algorithm and shows the comparison with k-mean algorithm. On the basis of comparison, the C-means algorithm is more accurate as well as efficient in speed than the K-means. Proposed c-means algorithm preserves the time. Performance of both c-means and k-means is demonstrated in the comparison outcome. Furthermore, in system we will implement classification algorithm over dimensional reduction dataset

REFERENCES

1. Christos Boutsidis, AnastasiosZouzias, Michael W. Mahoney, and PetrosDrineas, "Randomized Dimensionality Reduction for k-Means Clustering", IEEE Transaction on Information Theory, Vol. 61, NO. 2, FEBRUARY 2015.
2. Christos Boutsidis, Michael W. Mahoney, and PetrosDrineas, "Unsupervised feature selection for the k-means clustering problem," in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2009.
3. S. Arora, E. Hazan, and S. Kale, "A fast random sampling algorithm for sparsifying matrices," in Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (Lecture Notes in Computer Science), vol. 4110. Berlin, Germany: Springer-Verlag, 2006, pp. 272-279.
4. K. L. Clarkson, "Tighter bounds for random projections of manifolds," in Proc. 24th Annu. Symp. Comput. Geometry (SoCG), 2008, pp. 39-48.
5. S. Lloyd, "Least squares quantization in PCM," IEEE Transaction on Information Theory, vol. 28, no. 2, pp. 129-137, Mar. 1982.
6. I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror, "Result analysis of the NIPS 2003 feature selection challenge," in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2005.
7. PetrosDrineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, "Clustering in large graphs and matrices," in Proc. 10th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA), 1999, pp. 291-299.
8. PetrosDrineas, R. Kannan, and Michael Mahoney, "Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication," SIAM J. Comput., vol. 36, no. 1, pp. 132- 157, 2006.
9. S. Har-Peled and A. Kushal, "Smaller coresets for k-median and k-means clustering," in Proc. 21st Annu. Symp. Comput. Geometry (SoCG), 2005, pp. 126-134.
10. I. Kumar, Y. Sabharwal, and S. Sen, "A simple linear time (1 + ")- approximation algorithm for k-means clustering in any dimensions, in Proc. 45th Annu. IEEE Symp. Found. Comput. Sci. (FOCS), 2004, pp. 454-462.

BIOGRAPHY

Shilpa Mokashe is a ME Student of Department of Computer Engineering, RSCOE, Pune, Maharashtra, India.Prof. Sanjay B. Thakare, is the Assistant Professor of Department of Computer Engineering, RSCOE, Pune, Maharashtra, India.