International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 10 (2015) pp. 26845-26850

http://www.ripublication.com

# A Novel Centric Vector Based G-Means Algorithm For Text Dataset Clustering By Feature Selection Model

**Mr. Prabakaran M[1], Dr. C G Ravichandran[2]**
*Assistant Professor, Department of Computer Science and Engineering, RatnaVel Subramaniam College of Engineering and Technology, Dindigul, Tamilnadu, India[1]*
*Professor & Principal, Department of Computer Science and Engineering, SCAD Institute of Technology, Palladam, Tamilnadu, India[2]*
*prabakaranmtech@gmail.com[1], cg_ravi@yahoo.com[2]*

## Abstract

Data mining is an important domain for dataset clustering such as sentence clustering, since a sentence is likely to be related to more than one theme or topic present within a document or set of documents. It performs a better role in clustering process most sentence similarity measures do not represent sentences in a common metric space, conventional fuzzy clustering approaches based on prototypes or mixtures of centric vector are generally not applicable to sentence clustering. This paper presents a novel centric vector based G-means clustering algorithm that operates on relational input data; i.e., data in the form of a square matrix of pair wise similarities between data objects. In existing a novel fuzzy clustering algorithm that operates on relational input data. Results of applying the algorithm to sentence clustering tasks demonstrate that the algorithm is capable of identifying clusters of semantically related sentences, and that it is therefore of potential use in a variety of text mining tasks. Sentence clustering plays an important role in many text processing activities. In early classic summarization system, the important summaries were created according to the most frequent words in the text. The algorithm can be interpreted as a novel fast bottom-up Feature Selection Method for a clustering process.

**Key words:** Clustering, Document Organization, Vector Centric Feature Selection Model.

## Introduction

Data Mining is the process of extracting valuable information from the large amount of datasets, also popularly known as Knowledge Discovery in Databases (KDD), refers to the knowledge extraction of implicit, previously unknown and potentially

useful information from data in databases. It is a combination of some of the steps together. For instance, data cleaning, data integration and data transformation can be performed together as a data pre-processing phase to conduct KDD Process. Feature selection is a term commonly used in data mining to describe the methods and techniques available for reducing noisy data to extract some valuable informative data for pre processing and data analysis. Feature selection implies not only noisy reduction, which means imposing an arbitrary or predefined number of attributes that, can be considered when clustering a data, but also the number of attributes, which means that either the cluster points or grouping the similar labeled data selection attributes based on their vector centric.

**G-means algorithm**

The G-means algorithm is a combination of Gaussian model and K-means algorithm starts with a small number of k-means center positions which are called as centroid, and grows from the number of centric points[9]. The algorithm splits into two those center positions whose data points appear not to come from a Gaussian distribution. By using k-means clustering on the entire dataset and the entire data particles center positions to refine the current solution based on Gaussian distribution [8]. To initialize with just k = 1, or choose some larger value of k if we have some prior knowledge about the range of k[3].

Algorithm G-means( K,$C_{i,j}$)

1. Let C be the initial set of centers (usually C←{K(x)}).
2. C← K-means(C,X).
3. Let {class($x_i$) = j} be the set of data points assigned to center $C_j$ .
4. if each { $x_i$ |class($x_i$) = j} follow a Gaussian distribution(at best level).
5. If the data look Gaussian, keep $C_j$ . Otherwise replace $C_j$ with two centers.
6. Repeat from step 2 until no more center positions are added.

G-Means clustering methods are based on distribution set theory, and require that an object either does or does not belong to a center position. The Clustering idea is to cluster and classification performance is applied to analyze the accuracy of purity compare to sentence clustering.

**Feature Selection Model**

Feature selection is the process of selecting a subset of relevant features for use in model construction. It uses the mechanism of random selection on unlabeled data to form new training sets, and the most frequently selected feature is added to the result feature subset selection for each iteration. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features [4]. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Centric based Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points) [2]. Feature subset selection

is an effectual way for dimensionality reduction, elimination of inappropriate data, rising learning accurateness, and recovering result unambiguousness. Numerous feature subset selection methods have been planned and considered for machine learning applications [6]. They can be separated into four major categories such as: the Wrapper, Embedded, and Labeled and Hybrid methods. In order to evaluate the effectiveness of framework, give formal analysis as well as conduct extensive experimental study on the algorithm. The Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines [7]. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible.

**Proposed Architecture**

The basic idea of our proposed algorithm is from dataset clustering which using Gaussian distribution and K-means. The method ensures that there is exactly the requested number of clusters in the dataset. None of the existing distribution based techniques generate requested number based clusters. We proposed a method to generate such kind of clusters to improve the accuracy of clustered datasets. We also provide the clustering result along with the dataset, which can be used to verify the result of clustering algorithms.

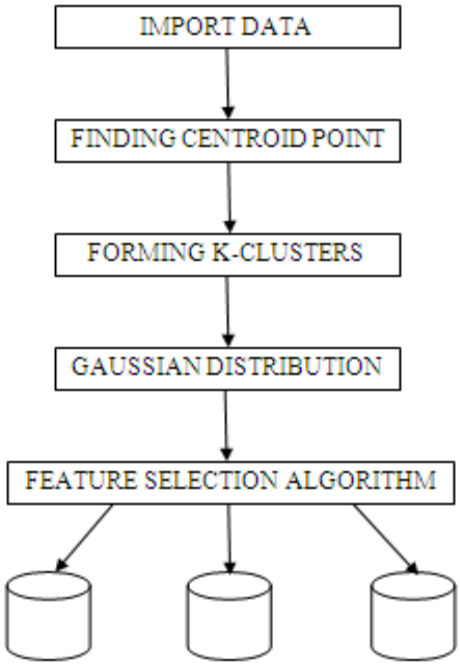


**Figure 1:** Architecture Diagram for Proposed Work

**Implementation Details & Experimental Results**

**BEGIN**
Initialize parameters;
Initialize cluster points randomly;
Initialize each centroid vector and velocity vector;
Find a velocity vector according to each text data particle's position;
Evaluate each particle and find the personal best and the global best;

**Do**
Update each data particle's velocity and position;
Find a sub set selection point according to the updated each particle's position;
Apply the local search;

**while (!Stop criterion) END**

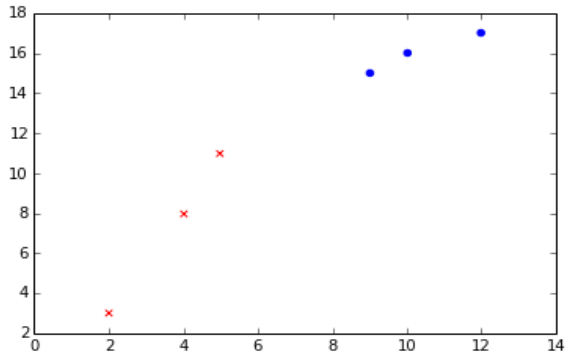


**Figure 2:** Five Centriod Points found by K-Means with number of iterations

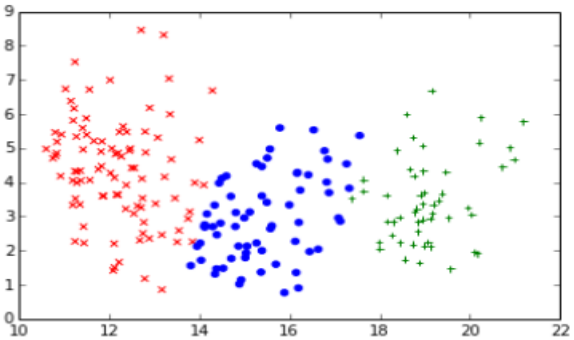


**Figure 3:** Huge number of Centriod Points found by G-Means with number of iterations

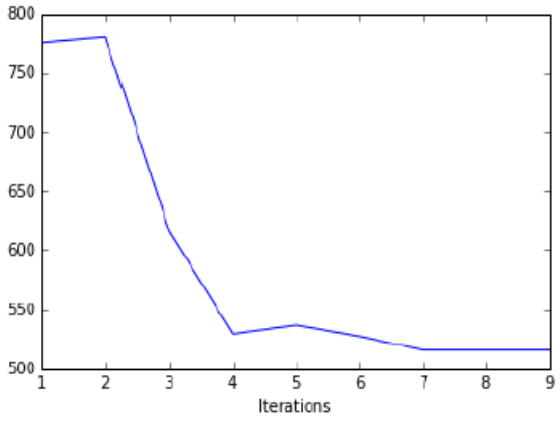


**Figure 4:** Iterations Vs Centroid Points

## Conclusion & Future Work

A good clustering of text dataset requires effective feature selection and a proper choice of the algorithm may perform a clustering task. Among the different classes of algorithms, the Gaussian-based clustering methods are among the most popular in a wide variety of applications. Based on this Gaussian based model, the clustering of labeled data can be improved. The experimental results show the implementation of the K means clustering and the Gaussian based feature selection model. In this study G-means has been implemented for a clustering process whereas we can implement other clustering algorithms e.g. Fuzzy C-means algorithm, EM algorithm and PSO Algorithm.

## References

[1] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988

[2] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, An Efficient k-Means Clustering Algorithm: Analysis and Implementation, July 2002

[3] Napoleon, D. and P.G. Lakshmi An efficient K-Means clustering algorithm for reducing time complexity using uniform distribution data points , 2010

[4] S. Z. Selim and M. A. Ismail, "K-means type algorithms: a generalized convergence theorem and characterization of local optimality," in IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 6, No. 1, pp. 81--87, 1984

[5] Jinzhu Hu, Chunxiu Xiong ,Jiangbo Shu,Xing Zhou,Jun Zhu, An Improved Text Clustering Method based on Hybrid Model, I.J.Modern Education and Computer Science, 2009

[6] YE Ping, Fuzzy K-means algorithms based on membership function improvement[J].Changchun Institute of Technology(Natural Sciences Edition), 2007

[7] Jian Zhu, Hanshi Wang, "An improved K-means clustering algorithm", The 2nd IEEE International Conference on Information Management and Engineering (ICIME), 2010

[8] Pritesh Vora and Bhavesh Oza "A Survey on K-mean Clustering and Particle Swarm Optimization", International Journal of Science and Modern Engineering Volume-1, Issue-3, February 2013

[9] Maria Isabel Ribeiro, Gaussian Probability Density Functions: Properties and Error Characterization, February 2004

[10] http://www.cs.cmu.edu/~awm/tutorials, Andrew W. Moore, Clustering with Gaussian Mixtures, 2004