A NEW FRAMEWORK FOR ENSEMBLING OF TEXT CLUSTERING DATA

[1]Dr. M. Ramakrishna Murthy, [2]Ch. Hemanth Krishna, [3]K Ananth Ramayya, [4]G Dinesh, [5]R Bhagya Sree

[1]Professor, [2]Student, [3]Student, [4]Student, [5]Student
Department of Computer Science and Engineering,
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India.

Abstract: A clustering ensemble algorithm aims to combine multiple clustering techniques to produce a better result than an individual clustering algorithm in this paper we describe a novel approach to clustering of text data. Text clustering is a technique through which text documents are divided into a particular number of groups so that the text documents within each group are related in content for these purposes we use two different clustering algorithms k-means and Birch Algorithm. Before using these algorithms, we perform the pre-processing of the documents Preprocessing techniques used are Stopword removal, pruning, stemming, Document represe- ntation-Vector Space model, after performing the preprocessing of the documents inverse document document frequency (IDF) has been achieved. These achieved IDF is used as an input to the clustering algorithms k-means and Birch.The common weighing scheme is TF-IDF (Term Frequency-Inverse Document Frequency),it has been found that the new weighting scheme word to vector provide better results than TF_IDF. We aim at applying the text clustering to articles like in newspaper using Word to Vector scheme to calculate the terms weight in the document vector. In this project the idea of ensemble text clustering of majority voting is used.

Keywords: **K-means,Birch, Word to Vector, Pre-processing.**

1. INTRODUCTION

Clustering is usually utilized in the world of pattern recognition and knowledge retrieval. Text clustering is a technique through which text documents are divided into a specific number of groups, in order that text within each group is same in contents. The goal of text clustering is to make a set containing relative data objects in a particular way like kind of text, a group of text, etc. In text clustering unsupervised technique of Text data is needed for clustering. Usually, text clustering techniques use characteristics like sequences and word phrases from the documents to apply the clustering. Text clustering is an interesting and advances research area due to the supply of an enormous amount of data in electronic forms. A lot of several applications are designed in literature which is applied for document clustering.

A large sort of clustering algorithms has been proposed: k-Means, EM (Expectation Maximization),hierarchical clustering algorithms like Single-Link, Fuzzy c-Means,etc. However, as it is known, there's no clustering method capable of correctly finding the underlying structure for all data sets So Clustering ensemble came into existence, combining different clustering results emerged as an alternative approach for improving the quality of the results of clustering algorithms.

Text Clustering with Word2vec

Computers cannot understand text data. We need to convert text into numerical vectors before any kind of text analysis like text clustering or classification Each word in word embeddings is represented by the vector. But let's say we are working with tweets from twitter and wish to understand how similar or dissimilar are tweets? So we'd like to possess vector representation of whole text in tweet to realize this we will do average word embeddings for every word in sentence (or tweet or paragraph).It does so in one among two ways, either using context to predict a target word (a method referred to as continuous bag of words, or CBOW), or employing a word to predict a target context, which is named skip-gram. In this paper we are going to discuss on CBOW(continuous bag of words)

Continuous Bag-of-Words Model

It Predicts center word from summing up the surrounding word vectors. These surrounding word vectors are referred to as the context of the center word.These proposed architecture is analogous to the feedforward NNLM, where it is the nonlinear hidden layer is removed and therefore the projection layer is shared for all words thus, all words get projected into an equivalent position (then the vectors are averaged). We call this architecture a Continuous bag-of-words model because the order of words within the history doesn't influence the projection. Furthermore, we also use words from the future; we've obtained the simplest performance on the task introduced within the next section by building a log-linear classifier with four future and four history words at the input, where the training criterion is to properly classify the present (middle) word. Training complexity is then

$$Q = N \times D + D \times \log2(V) \quad \text{..... (1)}$$

We denote this model further as CBOW, as unlike standard bag-of-words model, it uses continuous distributed representation of the context. The model architecture is shown at Figure1. Note that the weight matrix between the input and the projection layer is shared for all word positions in the same way as in the NNLM.

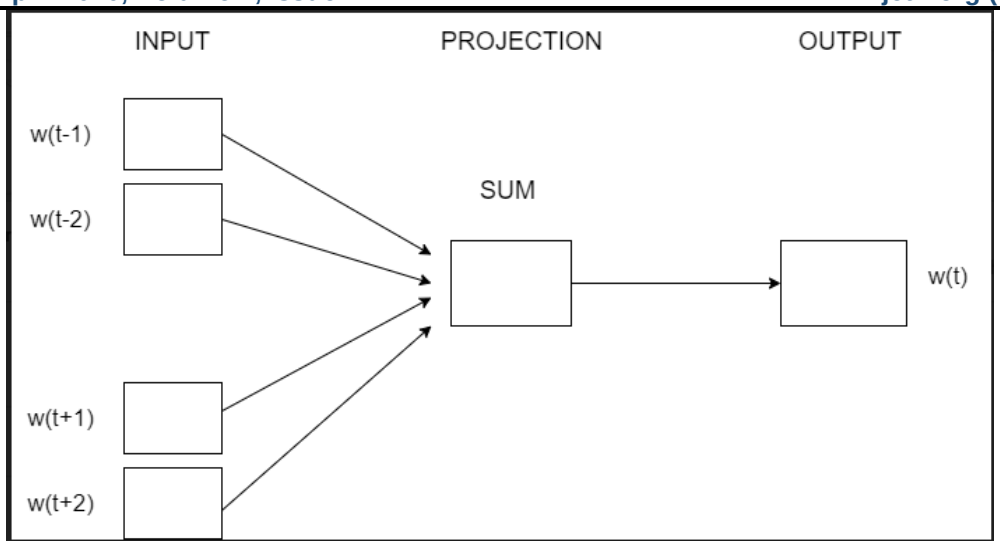


Fig 1 the cbow architecture predicts the current word based on the context

K-MEANS CLUSTERING

Clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data such that data points in the same groups are more almost like other data points within the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

In K-means clustering algorithm the no. of clusters required at the top need to be mentioned beforehand, which makes it important to possess prior knowledge of the dataset.

These models run iteratively to seek out the local optima.

If k is given, the K-means algorithm are often executed within the following steps:

Step 1: Partition of objects into k non-empty subsets
Step 2: Identifying the cluster centroids (mean point) of the current partition.
Step 3: Assigning each point to a specific cluster
Step 4: Compute the distances from each point and allot points to the cluster Where the distance from the centroid is minimum.
Step 5: After re-allocating the points, find the centroid of the new cluster formed.
Step 6: Repeat step 1 to step 6 until no improvements are possible.

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) CLUSTERING

A multiphase clustering algorithm (Zhang, Ramakrishnan & Livny, SIGMOD'96) Incrementally construct a CF (Clustering Feature) tree, a hierarchical arrangement for multiphase clustering.

Phase 1: Scan Text Documents to build an initial in-memory CF tree (a multi-level compression of the info that tries to preserve the inherent clustering structure of the data)

Phase 2: Use an arbitrary clustering algorithm to produce cluster of the leaf nodes of the CF-tree

Key idea: Multi-level clustering

Low-level micro-clustering: Reduce complexity and increase scalability. High-level macro-clustering: Leave enough flexibility for high-level clustering

Given a set of N d-dimensional data points, the clustering feature CF of the set is defined as the triple CF=(N, LS,,SS)
where

$LS \rightarrow = \sum_{i=1}^{N} \overrightarrow{x_i}$ is the linear sum of n points and

SS = $\sum_{i=1}^{N} (\overrightarrow{x_i})^2$ square sum of n points

Clustering feature:

Summary of the statistics for a given sub-cluster: the 0-th, 1st, and 2nd moments of the sub-cluster from the statistical point of view and registers crucial measurements for computing cluster and utilizes storage efficiently.

Centroid: $\overrightarrow{x_0}$

The "middle" of a cluster
n: number of points in a cluster
$\overrightarrow{x_i}$ is the i-th point in the cluster

$$C = \frac{\sum_{i=1}^{N} \vec{X}_i}{N} = \frac{\overrightarrow{LS}}{N} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Radius: R.

Average distance from member objects to the centroid
The root of average distance from any point of the cluster to its centroid
Average pairwise distance within a cluster
The root of average mean squared distance between all pairs of points in the cluster

$$R = \sqrt{\frac{\sum_{i=1}^{N} (\overrightarrow{x_i} - \vec{C})^2}{N}} = \sqrt{\frac{N.\vec{C}^2 + SS - 2.\overrightarrow{C}.\overrightarrow{LS}}{N}} = \sqrt{\frac{SS}{N}} - \sqrt{\frac{(\overrightarrow{LS}^2)}{(N)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

CF-Tree

Incremental insertion of new points (similar to B+-tree), For each point in the input, Find closest leaf entry and add point to leaf entry and update
CF{ If entry diameter > max diameter}
> split this leaf, and possibly parents > A CF tree has only two parameters
a. Branching factor: Maximum number of children
b. Maximum diameter of sun clusters stored at the leaf nodes

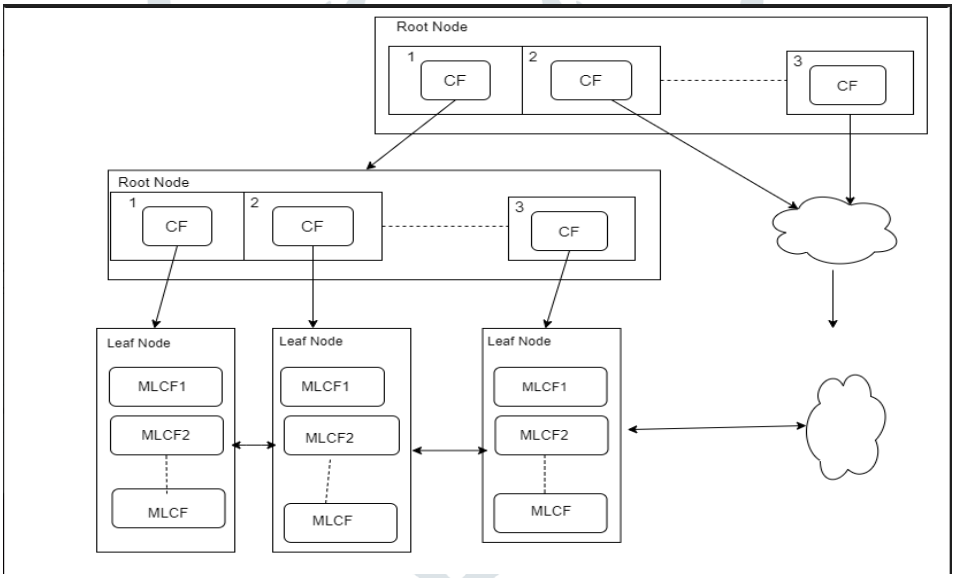


Fig 2 clustering feature tree

2. LITERATURE SURVEY

Sandro vega-pons and jose ruiz-shulclopery [1]. A large sort of clustering algorithms has been proposed: k-Means, Birch, supported spectral graph theory, hierarchical clustering algorithms like Single-Link, Fuzzy c-Means, etc. However, as it is known, there's no clustering method capable of correctly finding the underlying structure for all data sets. When we apply a clustering algorithm to a group of objects, it imposes an organization to the data following an indoor criterion, the characteristics of the used (dis)similarity function and therefore the dataset. Hence, if we've two different clustering algorithms and that we apply them to an equivalent dataset, we will obtain very different results.it supported the success of the combination of supervised classifiers. Given a group of objects, a cluster ensemble method consists of two principal steps: Generation, which is about the creation of a group of partitions of those objects, and Consensus Function, where a replacement partition, which is that the integration of all partitions obtained within the generation step, is computed.

Muhammad mateen [2] In this paper, clustering is discussed that's a group of comparable objects. Five clustering methods are applied on datasets, first on "Text clus" then on "20newsgroups" and obtained individual results. then an ensemble clustering technique is proposed supported major voting, to reinforce the performance of text clustering. During experiments on specified datasets, results of 5 clustering techniques represented within the above graphs, couldn't fulfill the wants , but ensemble clustering using majority voting technique proved fruitful for text clustering with better clustering results. Therefore, ensemble clustering is found better than five clustering techniques named as k-means, fuzzy c-means, agglomerative, k-medoid, and Gustafson Kessel. For future work, the ensemble clustering technique are often applied on text streams/web data clustering to separate contents and find the extremism content in it.

Tomas mikolov, Kai chen and Greg corrado, Jeffrey dean[3] The main goal of this paper is to introduce techniques which will be used for learning high-quality word vectors from huge data sets with billions of words, and with many words within the vocabulary. As far as we all know , none of the previously proposed architectures has been successfully trained on more than a couple of hundred of many words, with a modest dimensionality of the word vectors between 50 - 100.We use recently proposed techniques for measuring the standard of the resulting vector representations, with the expectation that not only will similar words tend to be on the brink of one another , but that words can have multiple degrees of similarity. This has been observed earlier within the context of inflectional languages for instance, nouns can have multiple word endings, and if we look for similar words during a subspace of the first vector space, it's possible to seek out words that have similar endings .Somewhat surprisingly, it had been found that similarity of word representations goes beyond simple syntactic regularities. employing a word offset technique where simple algebraic operations are performed on the word vectors, it had been shown for instance that vector(”King”) - vector(”Man”) + vector(”Woman”) leads to a vector that's closest to the vector representation of the word Queen .In this paper, we attempt to maximize accuracy othose vector operations by developing new model architectures that preserve the linear regularities among words.

Tahani Alqurashi, Wenjia Wang [4] A clustering ensemble aims to mix multiple clustering models to supply a far better result than that of the individual clustering algorithms in terms of consistency and quality. during this paper, we propose a clustering ensemble algorithm with a novel consensus function named Adaptive Clustering Ensemble. It employs two similarity measures, cluster similarity and a newly defined membership similarity, and works adaptively through three stages. the primary stage is to rework the initial clusters into a binary representation, and therefore the second is to aggregate the initial clusters that are most similar supported the cluster similarity measure between clusters. This iterates itself adaptively until the intended candidate clusters are produced. The third stage is to further refine the clusters by handling uncertain objects to supply an improved final clustering result with the specified number of clusters. Our proposed method is tested on various real-world benchmark datasets and its performance is compared with other state-of-the-art clustering ensemble methods, including the Co-association method and therefore the Meta-Clustering Algorithm. The experimental results indicate that on the average our method is more accurate and more efficient.

3.PROPOSED SYSTEM

The architecture flow is represented in the figure3. For the given input documents, for each document we perform preprocessing of the data containing 1. stop word removal 2. word to vector.

Given the text data(document) as input we first remove the stop words. Then the words are converted to vectors. Then we calculate the document similarly using cosine function. We then apply k means and Birch algorithms on the data.

The obtained data is used to implement. Comparative analysis is the stage where we apply real world inputs to know the running efficiency of the project.

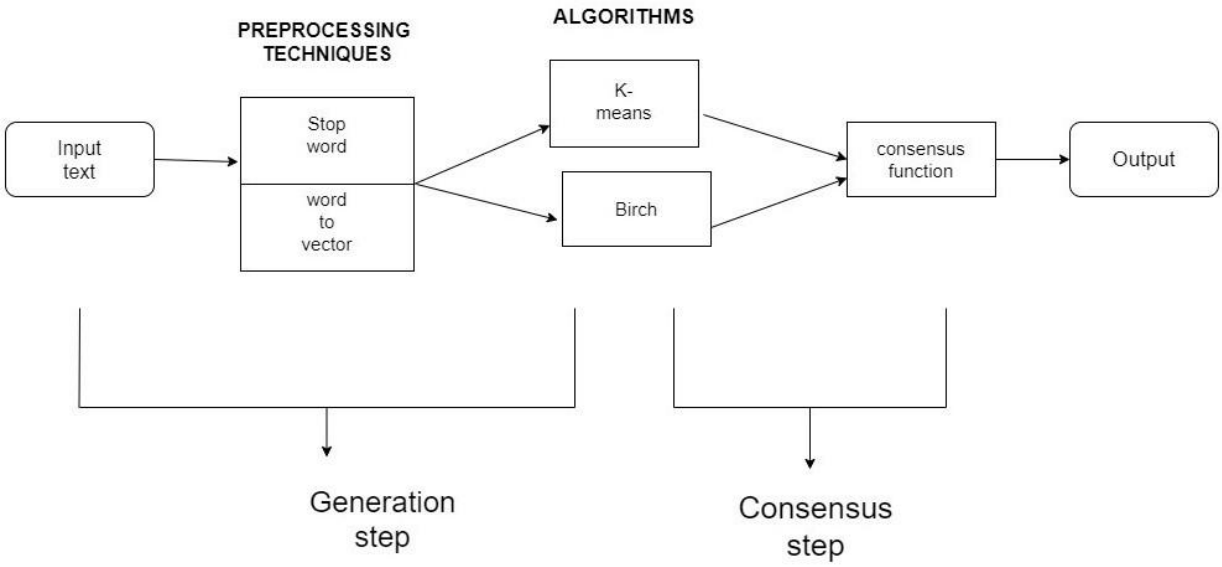


Fig 3 system architecture

Clustering Ensembles

Every Clustering Ensemble Technique is made up of two steps:

1 Generation Step
2 Consensus Step

Generation Step

The main aim here is to get m clustering models as the members for building the ensemble. In principle, any clustering algorithm might be used here as long as it is suitable for the dataset and it is important to use an appropriate generation process, because the final result are going to be conditioned by the initial clustering obtained during this step.

Preprocessing is used in the generation step before it is sent to the algorithms as input it involves transforming raw data into an understandable format. Data is said to be unclean if it is missing attribute, attribute values, contain noise or outliers and duplicate or wrong data. Presence of any of these will degrade quality of the results, so before we apply the clustering algorithms preprocessing techniques like stop word removal and word to vector is used. Data preprocessing is used by database-driven applications such as CRM and rule-based applications (like neural networks).

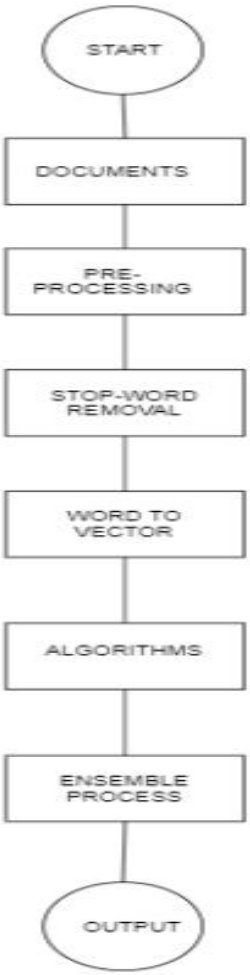


Fig 4 pre-processing process

In the generation step the weak clustering algorithms also are used. These algorithms structure a group of clustering using very simple and fast procedures. Despite the simplicity of this type of algorithms, Topchy et al showed that weak clustering algorithms are capable of manufacturing top quality consensus clustering in conjunction with a correct consensus function.

Consensus Step

A consensus step combines the output of different clustering algorithms {C1,C2,C3….Cm.} to obtain the final clustering result C*,Therefore it is considered as the most important step in clustering ensemble

A Consensus step has two approaches:

The object co-occurrence approach: It firstly computes the co-occurrence of objects within the members then determines their cluster labels to supply a consensus result. Simply, it counts the occurrence of an object in one cluster,or the occurrence of a pair of objects within the same cluster, and generates the ultimate clustering result by a voting process among the objects. Such methods are the Relabeling and Voting method the Co-association method and the Graph-based method

The median partition approach: Basic idea: Find a partition P that maximizes the similarity between P and everyone in the N partitions within the ensemble: P1, P2, …, PN and samples of this approach include the Non-Negative Matrix Factorization based method, the Genetic-based method and therefore the Kernel-based method .

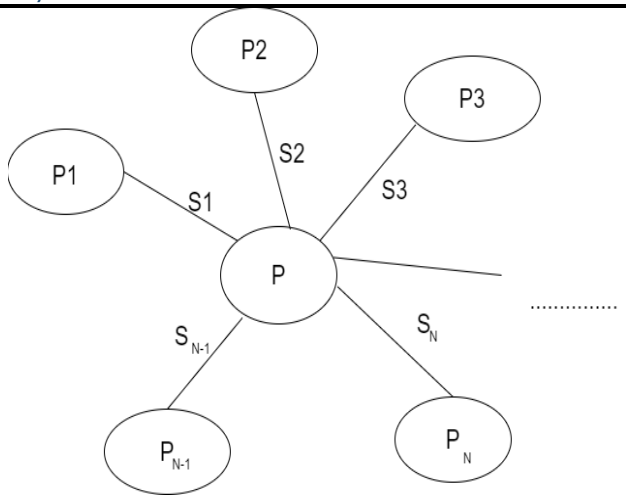
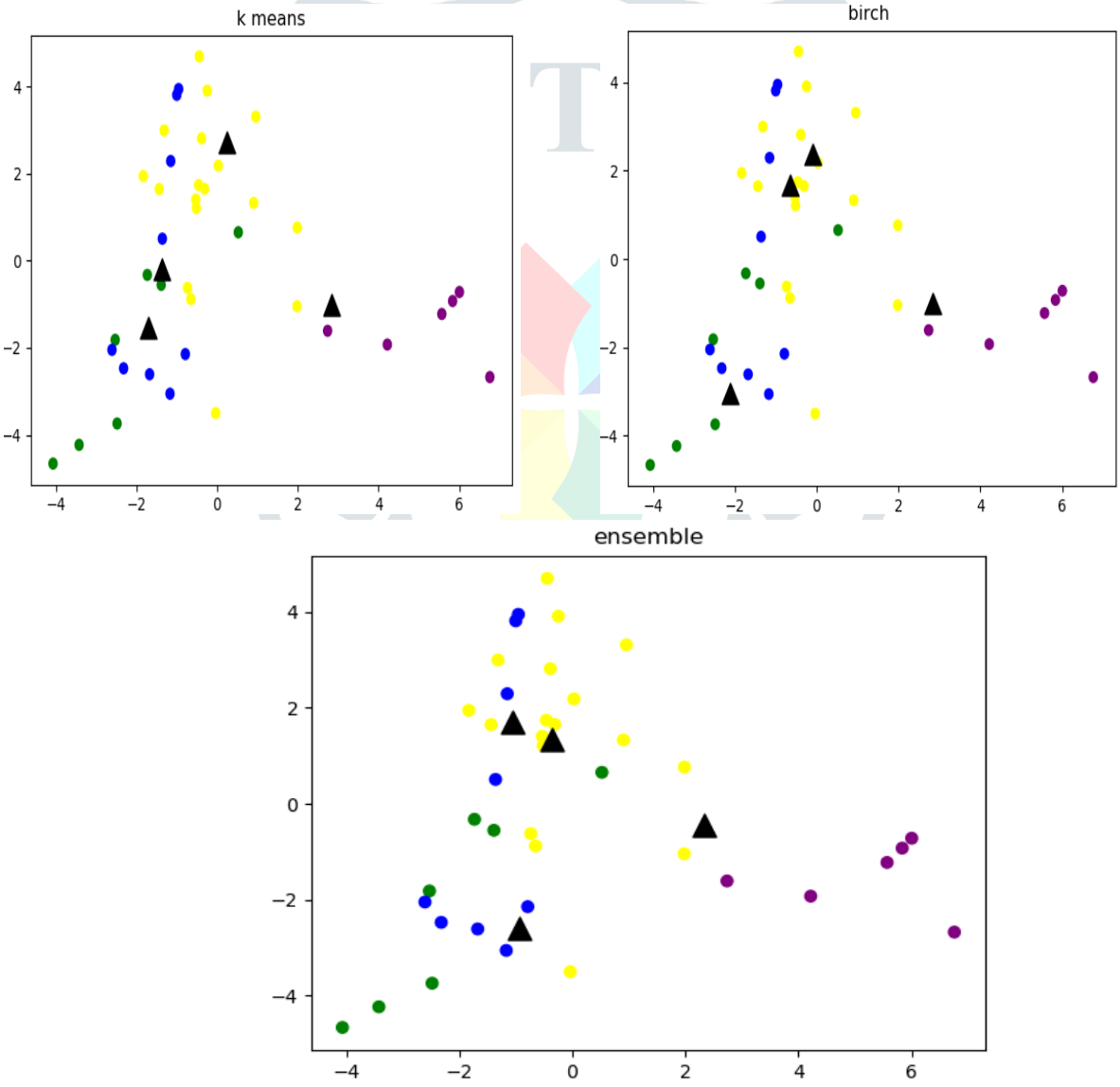


Fig 5 an example for median partition approach

4.EXPERIMENT ANALYSIS



INPUT: Text Documents

OUTPUT: Yellow dots represent cluster1, Green dots represent cluster2, Blue dots represents cluster3, Purple dots represent cluster4.

▲ Represents Cluster Centers

We did our project in python language, we used famous library matplotlib to plot and visualize all the results to matplotlib we gave input as vectors, centroids, labels which was converted into a 2-dimensional array using principal component analysis(this converts and n-dimensional array to m-dimensional array where m<n

However, these comparisons are made based on the experimental results obtained by applying different clustering algorithms to a fixed number of datasets ,Besides these comparisons are only among a few number of clustering ensemble methods(Relabeling and Voting, Co-association matrix etc.).In this sense, we should try to consider a more complete experimental comparison of clustering ensemble methods is necessary in order to give a benchmark results

5.CONCLUSION

Clustering ensemble has become a useful technique when facing cluster analysis problems, and its capacity for improving the results of simple clustering algorithms. The combination process integrates information from all partitions within the ensemble, where possible errors in simple clustering algorithms might be compensated. That way, the consensus clustering, obtained from a group of clustering of the same dataset, represents an appropriate solution. Due to the unsupervised nature of those techniques, it's not adequate to talk about the simplest clustering ensemble method.

Nevertheless, we can still establish a comparison among these methods and determine, for specific conditions, which one may be the most appropriate. We made analysis and comparison of the methods, taking under consideration different parameters. the most advantages of each method are often helpful to the users to pick the convenient method to resolve their problem.

6. REFERENCES

1. Sandro vega-pons and jose ruiz-shulclopery,"survey of clustering ensemble algorithmsInternational Journal of Pattern Recognition and Artificial Intelligence", May 2011

2. Muhammad mateen ,"Text Clustering using Ensemble Clustering Technique",International Journal of Advanced Computer Science and Applications, Vol. 9, No. 9, 2018

3. Tomas mikolov, Kai chen and Greg corrado, Jeffrey dean,"Efficient Estimation of Word Representations in Vector Space",arXiv:1301.3781v3 [cs.CL] 2013

4. Youguo Li, Haiyan Wu," A Clustering Method Based on K-Means Algorithm" ,© 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of [name organizer]

5. Usama M.Fayyad cory A.Reina Paul S.Bradley," Initialization of Iterative Refinement clustering algorithms"[C].Proc.4th International Conf.On Knowledge Discovery & Data Mining,1998.

6. JAIN A K, DUBES R C. "Algorithms for clustering data"[M].New Jersey:Prentice-Hall,1988.

7. Boris Lorbeer, Ana Kosareva , Bersant Deva , Dzenan Softi ˇ , Peter Ruppel , Axel Kupper ," Variations on the clustering algorithm BIRCH",Big Data Research 00 (2017) 1–11

8. Jashanjot Kaur,Preetpal Kaur Buttar,"A Systematic Review on Stopword Removal Algorithms",International Journal on Future Revolution in Computer Science & Communication Engineering ISSN: 2454-4248 Volume: 4 Issue: 4 207 – 210 ,April 2018

9. Parul Agarwal, M. Afshar Alam, Ranjit Biswal, "Issues, Challenges and Tools of Clustering Algorithms", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011 ISSN (Online): 1694-0814

10. Laith Mohammad Abualigah, Ahamad Tajudin Khader and Essam Said Hanandeh (2018), "Novel Weighting Scheme Applied to Improve the Text Document Clustering Techniques". Innovative Computing, Optimisation and Its Applications, Studies in Computational Intelligence 741,305-307.