Document Clustering: How to Measure Quality of Clusters in Absence of Ground Truth

[1]Iti Sharma, [2]Harish Sharma
[1]Dept of CSE, Career Point University
[2]Dept. of CSE, Rajasthan Technical University

Abstract
Demand of simple and scalable clustering algorithms for text documents is increasing as the volume of data generated by through internet is exploding. There are no known classes for such data and extrinsic measures of quality are not sufficient to guide about which algorithm is better for an application. This paper suggests four different intrinsic measures that can be used to evaluate cluster output and hence the clustering method to suit a particular application. The proposed metrics measure homogeneity and coherence of documents in a cluster as well as the overlap among different clusters in an interpretable form.

Keywords
Document Clustering, Spherical K-Means, Intrinsic Measures, Performance, Cluster Quality

I. Introduction

The increasing use of social media and Internet for newsfeeds has exponentially increased the amount of text documents available for repositories. Millions of web pages need to be indexed and organized for a fast search and retrieval system. Clustering is a preferred method to automatically organize similar documents together to club search results together or arrange documents in coherent topics. A popular algorithm to cluster text corpora is spherical k-means [1]. It is very flexible and many changes can be suggested to adapt it to suit requirements of application [2,3]. But there is no performance measure that can judge suitability of algorithm to clustering application. Existing extrinsic measures may not indicate algorithm that may be better suited to extract one most coherent topic out of many or to separate some overlapping topics.

The first indicator of performance of any clustering algorithm is the value of the objective function that is set for the clustering algorithm. Generally, the algorithms are tested on some popular or real-life data that has pre-associated cluster labels to each data object. The output of the clustering algorithm is cross-validated against the known labels. This leads to several possible techniques of evaluation of clustering quality through output labels. Such measures called extrinsic measures validate the clustering outcomes by already known or expected outcomes. Popular extrinsic measures are Accuracy, Precision, Recall, F-measure, Purity, Entropy, NMI, NCCG etc. Intrinsic measures on other hand measure the quality of cluster through the characteristics present in the cluster and not according to some known truth. They aim to measure the variations present among the members of the cluster so that homogeneity of a cluster can be established.

Clusters of text data cannot be solely evaluated on basis of extrinsic measures as the ground truth of text documents depends heavily on human perception. The amount of knowledge overlap among documents in any collection is high and all documents cannot be discretely categorized. So the class associated to them on basis of source or a human expert advice is always biased by perceptions. In this case, intrinsic measures are more useful to evaluate output of clustering algorithms for text documents.

Estivill-Castro [4] in his position paper argues that the reason why there are so many clustering algorithms is that notion of cluster cannot be precisely defined and depends largely on the perspective of the researchers. Therefore, if clustering algorithms are to be compared, the underlying inductive principle must be taken into account. Constructivist philosophy active scientific realism are used to argue that the idea of "truth" in cluster analysis depends on the context and the clustering aims. Different characteristics of clusterings are required in different situations [5]. Hence, relying totally on ground truth to validate clusters output by an algorithm is not justified.

This paper suggests different intrinsic measures that can be used to evaluate the output of a clustering algorithm for text documents.

II. Existing Performance Metrics

A. Precision, Recall, F-measure and Accuracy

We present the definitions as reported in [6]. Taking the case of only two cluster labels, the output labels could be true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn). Then the quality measures are computed as

$$Precision\ (P) = \frac{tp}{tp+fp}$$

$$Recall(R) = \frac{tp}{tp+fn}$$

$$F-score = \frac{(1+\beta^2)PR}{\beta^2 P + R}$$

Most commonly β is taken to be 1. The F-score can also be expressed in terms of true positives, false positives and false negatives:

$$F-score = s\frac{(1+\beta^2)tp}{(1+\beta^2)tp+\beta^2 fp+fn}$$

Accuracy is just the number of correctly classified points. In a binary situation,

$$accuracy = \frac{tp+tn}{N}$$

Where N is the total number of instances. For multiple classes accuracy is computed as follows:

$$accuracy = \frac{\sum_c^C tp_c}{N}$$

Where C is the number different cluster labels

For multi-class cases the precision, recall and F-measure can be computed as macro-averaged or micro-averaged. The general computations of these are taken from [7].

The macro-averaged results can be computed as indicated by: $L = \{\lambda_j : j=1\ldots q\}$ is the set of all labels. Consider a binary evaluation measure B(tp; tn; fp; fn) that is calculated based on. Let tp_λ, fp_λ, tn_λ and fn_λ be the number of true positives, false positives, true negatives and false negatives after binary evaluation for a label λ.

$$B_{macro} = \frac{1}{q}\sum_{\lambda=1}^{q} B(tp_\lambda, fp_\lambda, tn_\lambda\ fn_\lambda)$$

A micro-averaged can be computed as follows:

$$B_{micro} = B\left(\sum_{\lambda=1}^{q} tp_\lambda, \sum_{\lambda=1}^{q} fp_\lambda, \sum_{\lambda=1}^{q} fn_\lambda, \sum_{\lambda=1}^{q} tn_\lambda\right)$$

B. Purity and Entropy

These are also extrinsic measures. As reported in [8], let $\omega = \{w_1, w_2, \ldots, w_k\}$ be the set of clusters for the document collection D and $\xi = \{c_1, c_2, \ldots, c_j\}$ be the set of categories. Each cluster and category is a subset of the document collection, $\forall c \in \xi, w \in \omega : c, w \subset D$. Purity assigns a score based on the fraction of a cluster that is the majority category label,

$$\text{argmax}_{c\in\xi} \frac{|c \cap \omega_k|}{|\omega_k|}$$

in the interval [0, 1] where 0 is absence of purity and 1 is total purity.

Entropy defines a probability for each category and combines them to represent order within a cluster,

$$-\frac{1}{\log j}\sum_{j=1}^{j} \frac{|c \cap \omega_k|}{|\omega_k|} \log \frac{|c \cap \omega_k|}{|\omega_k|}$$

C. NCCG

This measure is proposed in [9]. The Cumulative Gain of a Cluster (CCG) is defined by the number of relevant documents in a cluster, $CCG(c, t) = \sum_{i=1}^{n} Rel_i$. A sorted vector CG is created for a clustering solution, c, and a topic, t, where each element represents the CCG of a cluster. It is normalized by the ideal gain vector,

$$\text{SplitScore}(t, c) = \sum^{|CG|} \frac{\text{cumsum}(CG)}{n_r{}^2}$$

where n_r is total number of relevant documents for the topic, t. The worst possible split places one relevant document in each cluster represented by the vector CG1,

$$\text{MinSplitScore}(t, c) = \sum^{|CG1|} \frac{\text{cumsum}(CG1)}{n_r{}^2}$$

NCCG is calculated using the previous functions,

$$\text{NCCG}(t, c) = \frac{\text{SplitScore}(t,c) - \text{MinSplitScore}(t,c)}{1 - \text{MinSplitScore}(t,c)}$$

D. Pair-based Indices

Another approach to define evaluation metrics for clustering is considering statistics over pairs of items [10]. Let SS be the number of pairs of items belonging to the same cluster and category; SD the number of pairs belonging to the same cluster and different category; DS the number of pairs belonging to different cluster and the same category, and DD the number of pairs belonging to different category and cluster. SS and DD are "good choices", and DS, SD are "bad choices".

Some of the Metrics using these figures are:

$$\text{Adjusted random Index R} = \frac{(SS + DD)}{SS + SD + DS + DD}$$

$$\text{Jaccard Coefficient J} = \frac{SS}{SS + SD + DS}$$

$$\text{Folkes and Mallows FM} = \sqrt{\frac{SS}{SS + SD}\ \frac{SS}{SS + Ds}}$$

III. A Synthetic Data Generator

Text data as in a corpus has directional characteristics, as has been observed by many including [1]. A very similar generative model is von Mises-Fisher distribution [11]. A d-dimensional unit random vector x (i.e., $x \in R^d$ and $\|x\| = 1$, or equivalently $x \in S^{d-1}$) is said to have d-variate von Mises-Fisher (vMF) distribution if its probability density function is given by

$$f(x|\mu, k) = c_d(k) e^{k\mu^T x}$$

where $\|\mu\| = 1$, $k \geq 0$ and $d \geq 2$. The normalizing constant $c_d(k)$ is given by

$$c_d(k) = \frac{k^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(k)}$$

where $I_r(\bullet)$ represents the modified Bessel function of the first kind and order r. The density f (x| μ,k) is parameterized by the mean direction μ, and the concentration parameter k, so-called because it characterizes how strongly the unit vectors drawn according to f (x| μ,k) are concentrated about the mean direction μ. Larger values of k imply stronger concentration about the mean direction. In particular when k = 0, f (x| μ,k) reduces to the uniform density on S^{d-1}), and as $k \to \infty$, f (x| μ,k) tends to a point density.

The vMF distribution is one of the simplest parametric distributions for directional data, and has properties analogous to those of the multi-variate Gaussian distribution for data in R^d. For example, the maximum entropy density on S^{d-1}, subject to the constraint that E[x] is fixed is a vMF density.

Hence, for a simple synthetic data generator (Fig. 1), we can use Gaussian distribution per term. Assuming to generate data from k different clusters, the number of documents of corpus is estimated not to be less than 4 times k, so that each cluster has at least 4 documents. Also, each document must have at least 10 terms, giving a lower bound on d as 10 times k. Now, few extra terms are required for the overlap. Each term is generated using a normal distribution for range [0,1] and mean 0.5. Thereafter, a constant α is subtracted from each and all negative values are replaced by zeros. This makes the columns sparse. For the columns pertaining to terms that are for overlap of clusters, a value of β is subtracted. Once this corpus is ready, each column is checked. The columns that have all zeros are marked and additional documents (rows) having non-zero values in these columns and few other randomly selected columns are constructed. This matrix is then normalized for the final output data matrix.

IV. Proposed Performance Metrics

Extrinsic measures are useful only when the ground truth is known or a gold standard extracted from human sense is available. In absence of these, as is the case of speedily generated text data of social sites and news feeds, such measures fail. Several researchers rely only on value of objective function in such cases.

Algorithm 1: Generate

//synthetic normalized document-term matrix generator

Input: number of clusters k, overlap γ, parameters α,β

Output: matrix X

n= 4*k

d= [10*k+2*n*γ]

Create sub-matrix of 4k rows and 10k columns as

$$X_{ij} \xleftarrow{\$,[0,1],0.5} \mathbb{R}^{4k \times 10k}$$

$X_{ij} = X_{ij} - \alpha$

For remaining columns, j= 10k+1 to d,

Pick randomly few indices *l* from 1 to 4k

Construct jth column

$$X_{lj} \xleftarrow{\$,[0,1],0.5} \mathbb{R}^{l \times 1}$$

$X_{ij} = X_{ij} - \beta$

Find columns p with all zero values, add γ rows to X, with random values in p columns

Output X

Fig. 1: Algorithm to Generate Synthetic Sparse Hyperspherical Data

We propose a few intrinsic measures that measure certain characteristics of the clusters such that they indicate quality of cluster.

A. Coherence and Adherence

We propose to measure the quality of clusters using two new matrices called adherence and coherence. Due to high dimensionality of text data and non-conformance to normal distribution, the cluster quality can be measured only through value of objective function. But objective function gives a collective quantity for entre partitioning. For a deep and rigorous analysis of each partition, we define more informative measures. A cluster may be compact if it has high similarity of member documents with each other and also has low similarity with documents of other clusters. This requires pair wise similarity computation among all document vectors making it computationally costly. Hence, we propose matrices that use quantities already being measured during clustering process and can be generated as a performance indicator along with output. No additional efforts for analysis need to be put.

Coherence as name suggests, measures the affinity of cluster member vectors towards the topic/concept vector (centroid). Every vector has some similarity with every concept vector, and maximum similarity with its assigned concept vector. First we count all the points in cluster that have high similarity with concept vector exclusively and similarity with any other concept vector rather low. This is possibly to test mathematically by comparing the similarity of a vector with all other centroids and see if it is lower than average. A vector that has high similarity only with its own centroid is called a coherent vector. Ratio of coherent vector members in a cluster is called coherence. Mathematically, this can be defined as

Definition 1: If $\mathbf{x}_i \mathbf{c}_j^T \leq \text{avg}_{j'=1\ldots k}\, \mathbf{x}_i \mathbf{c}_{j'}^T \;\; \forall j \in \{1, \ldots, k\}, j \neq y_i$ then $\mathbf{x}_i$ is a coherent vector.

Definition 2: The ratio of coherent vectors in a cluster to total number of members is coherence of the cluster.

$$Coh_k = \frac{\#\ coherent\ vectors\ in\ \pi_k}{|\pi_k|} \quad (1)$$

Thus, coherence measures the amount or proportion of documents in a cluster that are coherent with the topic denoted by the concept vector of that cluster and do not overlap much with other topics (clusters). Value of coherence for any cluster is between 0 and 1. Higher the value, more compact is the cluster.

Adherence is in contrast to coherence. While we expect clusters to be highly coherent, the mutual overlap of clusters should be least possible. This overlap or relation between two topics (centroids) is measured through adherence.

Definition 3: For any pair of centroids $\mathbf{c}_p$ and $\mathbf{c}_q$, adherence is the maximum similarity any point of $\boldsymbol{\pi}_p$ has with $\mathbf{c}_q$.

$$Adh_{pq} = \max_{\mathbf{x}_i \in \pi_p} \mathbf{x}_i \mathbf{c}_q^T \quad (2)$$

The boundary or edge between two clusters is formed of such points $\mathbf{x}_i$ that belong to one cluster but are sufficiently similar to centroid of other cluster. Adherence is an asymmetric relation between two clusters, that is it is not necessary to have equal value of Adh_{pq} and Adh_{qp}. Since Adherence is a similarity value, we can have its upper bound as 1 and lower bound as 0; but cannot have tighter bounds.

B. Density

An intuitive measure of density of a cluster can be as inspired from Physics. The number of vectors assigned to a cluster is its mass and density is its ratio to the volume of the cluster. The volume of the cluster is proportional to the angle that this hyper-arc subtends at the centre of hypersphere. Due to high dimensionality this becomes immeasurable. Hence, we come up with more suitable intuition for density. Consider a two-dimensional directional data with all points lying on circumference of a circle. Each cluster π_k (disjoint partition) is an arc of circle, subtending some angle φ_k at centre. This angle is actually difference of the maximum angle from x-axis of any member of π_k and minimum angle from x-axis of any member of π_k.

$\varphi_k = \varphi_{k\,max} - \varphi_{k\,min}$

Hence, density could be $|\pi_k| / \varphi_k$, but observing the difficulty of measuring φ_k, we see that the angle is proportional to size of arc and chord, and size of chord is proportional to size of x-interval spanned by the cluster. Instead of taking x-axis, in multi-dimensional vector space of text documents, we propose a principal dimension. And measure size of principal dimension spanned by a cluster as its volume. Accordingly, density is computed.

Definition 4: The principal dimension of any corpus corresponds to the term having highest normalized frequency in the corpus.

$$\alpha = \underset{j}{\text{argmax}} \sum_{i=1}^{n} x_{ij}$$

Definition 5: Density of a cluster is ratio of its population to the length of principal dimension spanned by the cluster.

$$\rho_k = \frac{|\pi_k|}{\max_{x_i \in \pi_k} x_{i\alpha} - \min_{x_i \in \pi_k} x_{i\alpha} + 1} \quad (3)$$

Thus, density is also a measure of compactness of a cluster. If a cluster contains documents belonging to variety of subtopics, the range spanned by the cluster is large and density will be low indicating the diversity of cluster. If a cluster contains documents related to single or limited set of topics, the span is lesser and density higher. The value of 1 is used as a smoothing factor to have singleton clusters density =1 and avoid division by zero situation.

C. Inter-cluster Separation

This is an extrinsic measure as it measures the separation of clusters. The clusters may not be separated inherently, but to understand underlying patterns, the clustering method always tries to identify the groups which are very different from each other. The centroid vector represents the mean direction of all cluster member documents. If similarity between these centroids is measured, it would give the angle between centroids. Since cosine of an angle is inversely proportional, larger similarity value means smaller angle. To give a direct interpretation, we take separation as difference of similarity to itself and similarity to another centroid.

Definition 6: The inter-cluster separation is the distance between the centroids of the clusters.

$$ICS_{pq} = \mathbf{c}_p \mathbf{c}_p^T - \mathbf{c}_p \mathbf{c}_q^T \quad (4)$$

V. Experimental Evaluation

The synthetic generator is used to generate several datasets at different parameter settings and relation between suggested performance metrics and established metrics like ARI and objective function of the SKM is evaluated by plotting them together. At k=5, overlap γ=0.3, parameters α=0.4,β=0.3, dataset is generated and the SKM is run for 50 times. The values of objective function, density and maximum adherence is adjusted to fit in scale to show the plots as in fig .2, 3 and 4. It is observed in fig. 2 that the pattern of ARI is not followed by objective function or average similarity of largest cluster while these both have very close similarity to each other. This can be interpreted as that ground truth may not conform to the best value of objective function. In fig. 3 the ARI is plotted with coherence of largest cluster and density of largest cluster. Both of these appear to have more close growth pattern as of ARI. Thus, these proposed intrinsic measures can be used when ARI cannot be measured due to absence of ground truth. In fig. 4 maximum and minimum adherence values between the clusters is measured for each run and plotted with respective values of ARI. The dissociation among these measures is very clear and indicates that adherence measure can give a different insight into the cluster structure very different from ARI or any accuracy. Adherence is a measure of overall output structure and does not correspond to any individual cluster or existing class in the data.

VI. Conclusion

Clustering of text data is required for many tasks that are becoming more relevant in today's scenario of growing social media. Spherical k-means is a good linear time algorithm for this purpose that has much flexibility to be adapted according to requirements of application. But the measures that can guide about the algorithm's suitability to an application are not available. Several extrinsic measures are available but ground truth is not known in case of social media documents or many news feeds that belong to overlapping classes. This paper suggests four intrinsic performance measures to evaluate quality of output of clustering algorithm for text documents. Density describes compactness of a text cluster; coherence describes how many documents purely pertain to topic without any overlap; adherence describes the overlap of topics among clusters as it is a pairwise metric and can also serve as an objective if bounds on minimum or maximum adherence are provided.

References

[1] I.S. Dhillon, D. S. Modha,"Concept decompositions for large sparse text data using clustering", Machine Learning, Vol. 42, No. 1, pp. 143–175, 2001.

[2] I.S. Dhillon, J. Fan, Y. Guan,"Efficient clustering of very large document collections", In: Data Mining for Scientific and Engineering Applications, Norwell, MA: Kluwer Academic Publishers, pp. 357–381, 2001.

[3] K. Hornik, I. Feinerer, M. Kober, C. Buchta,"Spherical k-means clustering", Journal of statistical software, vol. 50(10), pp. 1-22, 2012.

[4] V. Estivill-Castro,"Why so many clustering algorithms: a position paper", ACM SIGKDD Explorations Newsletter,vol. 4(1), pp. 65–75, 2002.

[5] C. Hennig,"What are the true clusters?", Pattern Recognition Letters, Vol. 64, pp. 53-62, 2015.

[6] V. Van-Asch (2013),"Macro- and micro-averaged evaluation measures." Belgium: University of Antwerp, [Online] Available: https://www.semanticscholar.org

[7] G. Tsoumakas, I. Katakis, I. P. Vlahavas,"Mining multi-labeldata", In Data Mining and Knowledge Discovery Handbook, 2nd ed., Part 6, O. Maimon, & L. Rokach (Ed.). Heidelberg, Germany: Springer-Verlag, pp. 667-685, 2010.

[8] C.M. De Vries, S. Geva, A. Trotman,"Document clustering evaluation: Divergence from a random baseline", arXiv preprint arXiv:1208.5654

[9] R. Nayak, C.M. De Vries, S. Kutty, S. Geva, L. Denoyer, P. Gallinari,"Overview of the INEX 2009 XML mining track: Clustering and classification of XML documents", In proceedings of 8th International Workshop of the initiative for the Evaluation of XML Retrieval, INEX 2009, Brisbane, Queensland, pp. 366–378, 2009.

[10] E. Amigo, J. Gonzalo, J. Artiles, F. Verdejo,"A comparison of extrinsic clustering evaluation metrics based on formal constraints," Journal of Information retrieval, Vol. 12(4), pp. 461-486, 2009.

[11] A Banerjee, J Ghosh,"Frequency-sensitive competitive learning for scalable balanced clustering on high-dimensionalhyperspheres", IEEE Trans. Neural Networks, Vol. 15, No. 3, pp. 702–719, 2004.

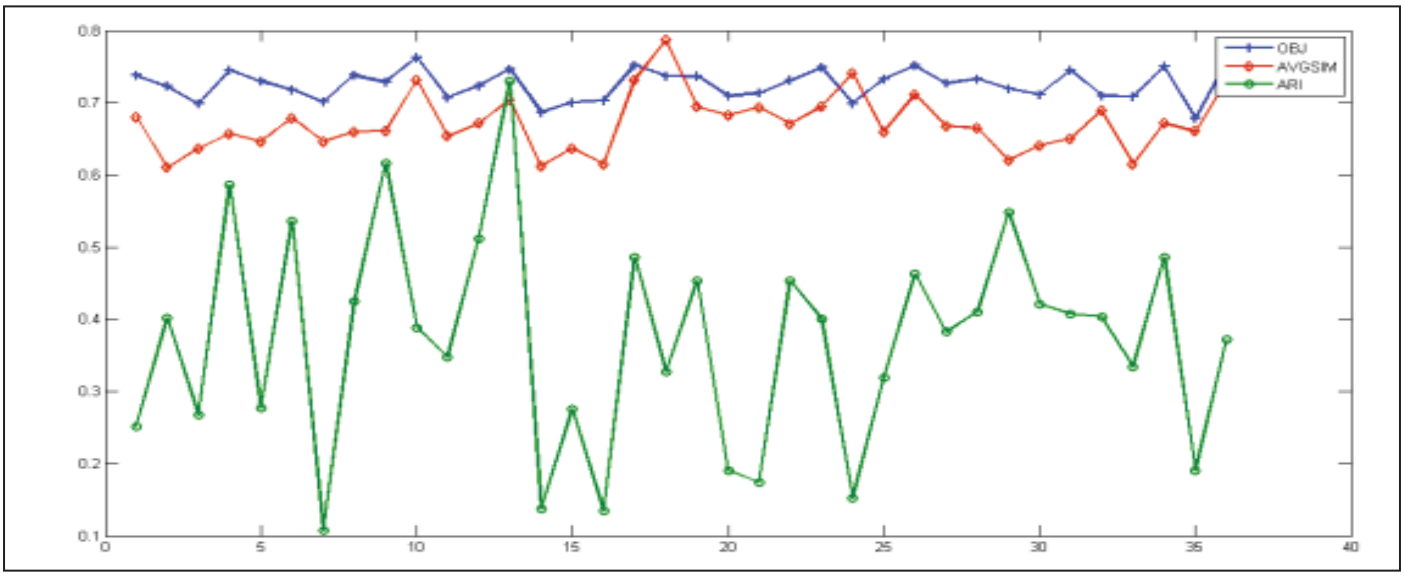

Fig. 2: Variation in objective function value and average similarity of documents to centroid in largest cluster as compared to variation in ARI

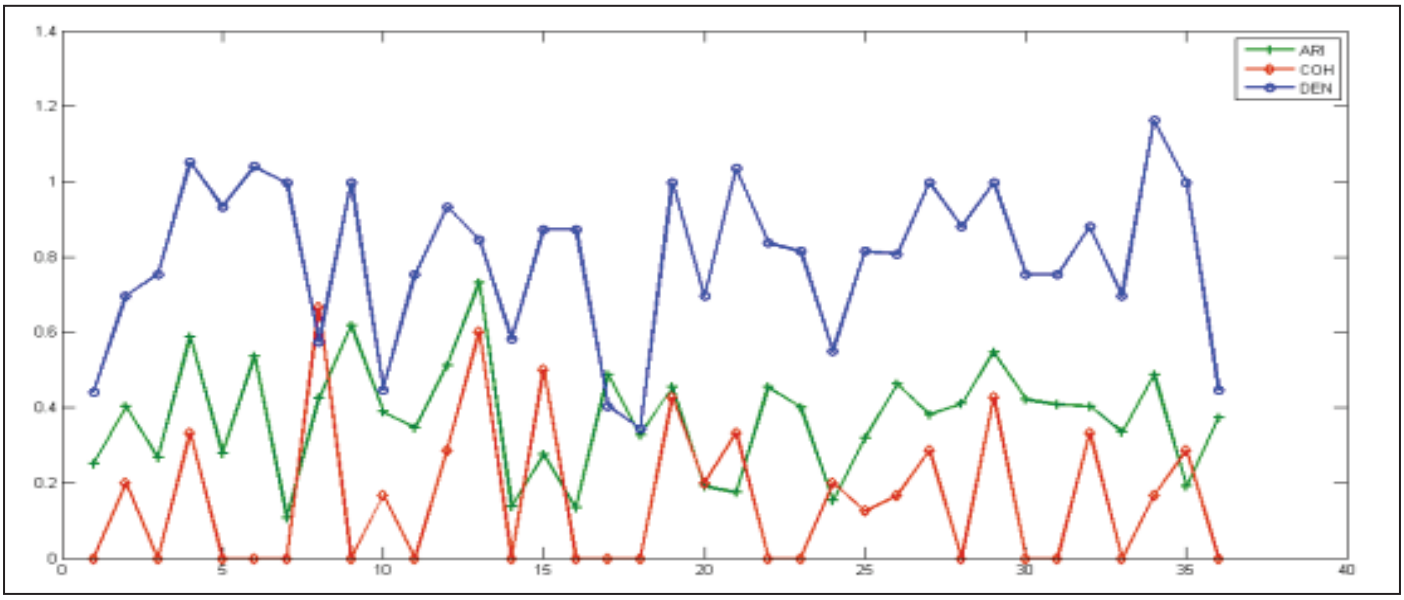

Fig. 3: Variation in coherence and density of largest cluster as compared to variation in ARI

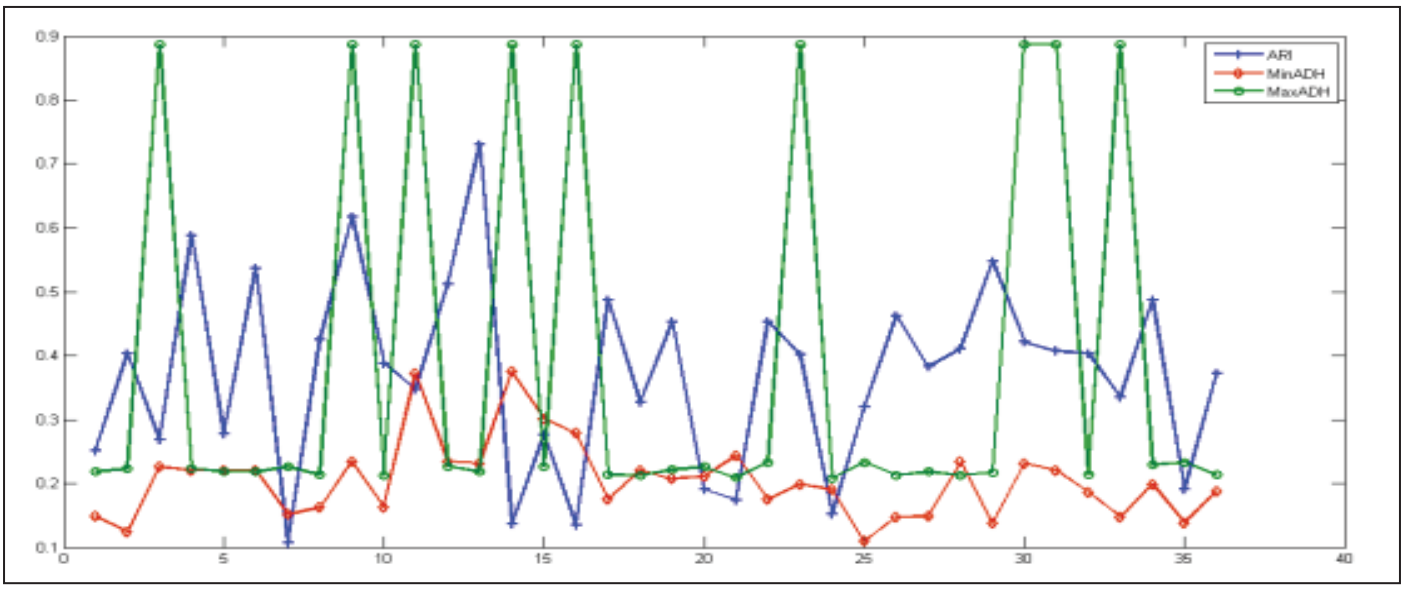

Fig. 4: Variation in maximum adherence and minimum adherence among the clusters as compared to variation in ARI