Assignment on Fuzzy Clustering
and
Topic Modelling

In partial fulfillment of course on Advanced Data Mining

BITS Pilani

Submitted on date: 04/03/2016

Team members

Sanket Shah - 2013A7PS119P
Akhilesh Sudhakar - 2013A7PS173P

Paper used:

Efficient Stochastic Algorithms for Document Clustering (Forsatia, Mahdavi, Shamsfarda, Mohammad Reza Meybodi, 2013)

Aim:

To cluster a group of documents such that each document belongs to all clusters, but to varying degrees (soft-clustering), and to retrieve documents from a topic, given keywords belonging to that topic.

Dataset:

Times of India articles from January 2016 in the topics: 'Lifestyle', 'Business', 'Entertainment', 'Sports', 'Tech'.

I**nput**:

A corpus of documents, number of clusters.

Output:

1. Clustering: Cluster centres, belonging of each document to cluster centers.
2. IR System: Takes as input a phrase/keyword/document and gives as output similar documents.

Flow of Control:

1. Pre-processing: Removing stop-words, performing stemming/lemmatization, converting documents from words to tf-idf vectors
2. Harmony Search:
3. k-Means/FCM:
4. Evaluation of clustering results
5. IR system:

Improvements:

1. Objective function:

The algorithm (Harmony Search) uses an objective function called ADDC minimization (Average Distance of Documents to Cluster centers) using the distance metric as Euclidean distance.

-- An improvement made to this was to also adapt the Davies-Bouldin Index to be adaptively used as a metric of clustering quality. The solution corresponding to the minimum value of the DB index is considered the best solution.

--Further, cosine similarity is also used as a similarity metric for documents with their cluster centers. The document corresponding to the maximum cosine similarity with its cluster center is considered the best.

2. Fuzzifying the hybrid algorithm: (HSCLUST) + K-Means :

The paper talks about using HSCLUST to build an initial clsuter configuration for all the documents, but in a non-fuzzy manner. The improvement made has been to adapt this to the domain of fuzzy clustering of documents, using fuzzy C-means approach.

IR System:

A query system accepts as input from the user, an arbitary number of keywords. Further retrieval happens by identifying topics that the given set of keywords could belong to.

Results:

As the clustering of the documents had to be evaluated, the topics returned by the clustering were compared with the topic labels of the data. The ground truth topic labels were: 'Lifestyle', 'Business', 'Entertainment', 'Sports', 'Tech'. In order to compare the topics obtained vs the pre-known topics, the Rand Index was used. The Omega Index could not be used as pre-known fuzzy labels are not available for documents.

Improvement Combinations	**RandIndex**
K-Means/Stem/ADDC	0.947750865
K-Means/Stem/DB	0.947750865
K-Means/Lemmatize/ADDC	0.947750865
K-Means/Lemmatize/DB	0.925190311

FCM/Stem/ADDC	0.736401384
FCM/Stem/DB	0.732110727
FCM/Lemmatize/ADDC	0.723529412
FCM/Lemmatize/DB	0.654809689

Conclusions:

The following combinations were used while simulating results:

(Stemming vs lemmatizing) X (Cosine similarity ADDC vs Davies Bouldin index) X (fuzzy and non-fuzzy clusters)

Hence, a total of 8 rand index values were obtained for each run of the clustering. Moreover, we also noticed that these 8 values did not change much across different runs, and hence concluded that even though the initial assignment of cluster centers was stochastic, the clustering algorithms made it predictable.

All results were measured using the rand index. The best results were obtained using Harmony Search to find the initial configuration of clusters by using ADDC of cosine similarity and then applying K-means on this and fuzzifying it later. Not too much variance in the value of the rand index was obtained across different combinations.