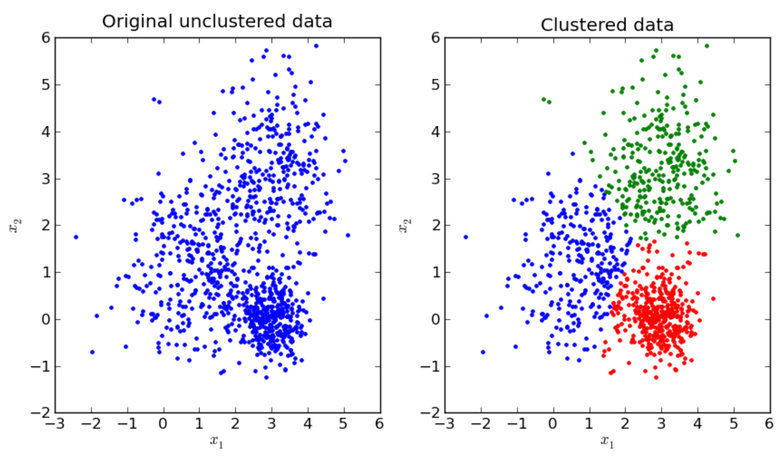
Comparison of Clustering Algorithms in Text Clustering Tasks

Rafael Gallardo-García, Beatriz Beltrán, Darnes Vilariño, Claudia Zepeda, Rodolfo Martínez
Benemérita Universidad Autónoma de Puebla, Mexico
rafael.gallardo@alumno.buap.mx,
{bbeltran,darnes,czepedac,beetho}@cs.buap.mx

Language & Knowledge Engineering

Introduction

The task of grouping a set of objects in a cluster, where the objects inside are more similar to each other in the same cluster than objects in other clusters is called *Cluster Analysis* or *Clustering*, this is one of the most important research areas in present.
It is presented a comparison between three partitional clustering algorithms: *K-Means*, *Affinity Propagation* and *Spectral Clustering*, there algorithms were specifically used in *document clustering* and *text clustering* tasks (some of the most important research areas in NLP).



How partitional algorithms work?

In this kind of algorithms, clusters are determined promptly, the partitioning algorithms divide the data into a partitions, where each partition represent a cluster. In this algorithms each cluster must contain at least one object, and each object must belong to exactly one group.

Affinity Propagation

Affinity Propagation forms clusters by sending messages between pairs of instances until convergence and considers all data points as potential exemplars.
This algorithm works by sending messages between a "Responsability Matrix" and an "Availability Matrix", these matrices are initialized with zeros, then the algorithm realize the next updates iteratively:

$$r(i,k) \leftarrow s(i,k) - \max_{k \neq k'} \{a(i,k') + s(i,k')\}, \quad (1)$$

$$a(i,k) \leftarrow \min\left(0, r(k,k) + \sum_{i' \notin \{i,k\}} \max(0, r(i',k))\right) \; for\, i \neq k, \quad (2)$$

$$a(k,k) \leftarrow \sum_{i' \neq k} \max(0, r(i',k)). \quad (3)$$

(1) is where responsability matrix updated are sent around and (2), (3) are the way to update the availability matrix.
Once the cluster boundaries stop changing over a number of iterations, the algorithm extracts the exemplars and clusters from the final matrices as those whose "responsibility + availability" be positive, (i.e (*(r(i, i) + a(i, i)) > 0*).

K-Means

K-Means uses vector quantization methods in order to get a partitioning of the data space into Voronoi cells returned as clusters. K-Means commonly uses a random initial points called "seeds" and then proceeds by alternating between two steps to select new cluster centroids.
The partitioning of the observations according to the Voronoi diagram generated by the means, is calculated as:

$$S_i^{(t)} = \{x_p : ||x_p - m_i^{(t)}||^2 \leq ||x_p - m_j^{(t)}||^2 \forall j, 1 \leq j \leq k\}, \quad (4)$$

Calculating the new centroids within the new clusters is realized with the following equation:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|}. \quad (5)$$

The algorithm has converged when assignments stop changing, that means what the Voronoi cells (clusters) are complete.

Spectral Clustering

Spectral Clustering algorithm uses the eigenvalues of a similarity matrix obtained from the data, to realize dimensionality reduction and then it makes clustering in less dimensions.
Spectral Clustering works on relevant eigenvector of a Laplacian Matrix of A, this algorithm project the data into a lower-dimensional space (the eigenvector domain) where they are easily separable with other methods, once the dimensionlity reduction is complete, the process for cluster the data in this lower-dimensional space is the same that K-Means.

About the tests

The supervised corpus was taken from the PAN , which consists in 60 problems, each problem contains 20 texts and have the gold standard. Before try to cluster the documents, we got a mathematical form of each text using *tf-idf,* and we build the similarity matrixes by using *cosine similarity.* Each algorithm's set of clusters will be compared with the supervised corpus, accuracy averages are in *f-measure* terms (the best results were taken).

Results

The following table shows the *precision, recall* and *f-measure* averages of the best test results.

Algorithm	Precision	Recall	F-measure
Affinity Propagation	**0.704**	0.606	0.651
Spectral Clustering	0.694	**0.833**	**0.758**
K-Means	0.619	0.747	0.677