Boosting LVM-based Document Clustering and Visualization with Genetic Chromodynamics

Abstract

We present probabilistical models for text document analysis and visualization by means of clustering and topographic organization. These models are capable of revealing the underlying semantic structure in text based documents, and facilitate human interpretation of the high dimensional data by two dimensional visualization.

However, results of applying *latent variable models* are under the influence of initialization values, and setting the number of latent variables beforehand is necessary. To solve these problems, we have resorted to *genetical chromodynamics*, a new and potentially appealing strategy for multimodal optimization and search problems.

To illustrate these models and examine their efficiency, we developed an application that solves two text mining problems: unsupervised classification and 2D visualization of text documents.

In order to discover the latent structure of the observable data without supervision, we apply *generative models*, frequently used in areas like natural language processing, data mining, fraud detection, information retrieval, bio-informatics. The *latent class model* provides probabilistical labels (i.e. the assignment of a document to a cluster is soft, not categorical) useful for the interpretation of the results. These are highlighted by two dimensional visualization, achieved by applying the *latent trait model*.

Given a text corpus, documents are likely to include terms highly indicative of one or relatively few topics, together with noise terms. Statistical pattern recognition and information retrieval algorithms are built on the premise that the patterns we observe are generated by random processes that follow specific distributions. Specifically, we make the assumption that a distribution of the words in the dictionary is assigned to each cluster, respectively grid point. Proposing credible distributions that can generate natural language is very difficult, and the computation involved is usually heavy-duty. Therefore, we must be satisfied to model only a few aspects of the observed data. For example, we won't consider dependencies and ordering between terms, and we assume that term occurrences are independent events.

We propose generative models of documents that don't define similarity or distance measures: we assume random processes, called *latent variables* generating the documents. Our goal is the discovery of the variables that are most likely to have generated a given collection of documents and the associated parameters. In the case of text document clustering, the latent variable is played by topic categories. The latent dimension is the number of clusters, previously given. On the other hand, the starting point for the latent

trait variables will be a two dimensional uniform grid of points, mapped by a set of non-linear and linear basis functions to the latent space (the latent dimension is previously given in this case, too). The basis functions can be arbitrarily chosen among smooth functions – in our application, we used radial basis functions with constant unit variance.

For the representation of data in such a form that facilitates actual processing, we choose the *vector space model*: documents are represented as vectors in a multidimensional Euclidean space, each axis in this space corresponding to a term (token). Naturally, all different words in all documents should be taken into consideration, but the usage of synonyms and different writing styles excessively increases the dimensionality of the word space. This is obviously inconvenient as implies the manipulation of an extremely high dimensional data matrix; therefore, we use a mindfully selected set of words (the size of this dictionary is previously fixed).

In order to determine the parameters of the models from the data set, we use the *Expectation Maximization (EM) algorithm*. This algorithm is based on maximizing the data likelihood: iteratively modifies the parameters to decrease the error function, thus increasing the likelihood. This is preceeded by randomly setting the parameter values. It's important to mention that initial values of the parameters greately influence which local minima is found, and thus the results of the whole clustering process.

During our experiments, we observed that these models have found valid meaningful mappings, as related document instances were clumped together on the maps, respectively into clusters, while those concerning different topics were mapped further from each other, respectively grouped into separate clusters. Thus, we can draw the conclusion that the latent class and trait models are suitable for topographical visualization and semantic structure discovery in corpora containing text documents. Note that the assumption of data instances being independent and identically distrtributed was made. These methods are useful for any such data analysis applications where the visualization and clustering of high dimensional possibly sparse multivariate discrete data set is sought.

Clustering requires the number of clusters to be given, i.e. how many latent variables do we suppose to have generated the text documents. However, we can set this value arbitrarily, the process resulting in more or less refined clusters. On the other hand, clustering results greatly depend on initialization values i.e. clustering the same dataset over and over, we may obtain different classifications although the number of clusters remains the same. Originally, we used *k-means clustering* to obtain initial cluster centers. To tackle both problems (find the optimal number of clusters and make clustering more consistent), we experimented with *genetic chromodynamics* to see if it is a viable solution. Standard genetic algorithms fail to detect multiple optimum points, on the other hand premature local convergence does present difficulty. The strategy we have considered is a flexible method intended to solve multimodal optimization and search problems in distributed artificial intelligence applications, like cooperative multi-agents. The main idea of this approach is to force the formation and maintenance of sub-populations of solutions. These sub-populations co-evolve and eventually converge towards several local and global optimal solutions. We examined whether genetic chromodynamics automatically provides cluster centers that meet our requirements.