Unsupervised Document Classification using Sequential Information Maximization

N. Slonim et al

Summary by M. Pollari

In the article Slonim *et al* presented an Information Bottleneck (IB) framework for unsupervised document clustering. Unsupervised document clustering is needed for retrieving information and/or navigating and browsing large document collections. The idea is to find clusters from collection of unlabeled documents that correspondence the true topics of the documents.

It has been shown that IB based methods are well suited for unsupervised document clustering. However, most of the earlier IB methods have utilized agglomerative procedure, which have two major drawbacks: There is no guarantee that agglomerative clustering can find a solution that is local maximum of the target function. Furthermore they are computationally expensive and thus infeasible for large document collections. In this article a sequential Information Bottleneck (sIB) was proposed to overcome these limitations.

Most of the clustering algorithms are based on distance or distortion measures. The selection of the distance (distortion) measure is, however, often arbitrary. A natural measure of similarity between documents is their word conditional probabilities. For instance if words; *document, clustering, Information, and Bottleneck*, have high probabilities in two documents then most likely these two documents are from same topic and should belong to same cluster. This is known as conditional clustering i.e. we a looking a cluster hierarchy for a one set(documents) based on the similarity of their conditional distributions with respect to another set (words). However, one has not solved the "right" distance measure yet because the are still infinitely many possible distance measures which use the same principle. The Information Bottleneck principle [Thisby et al.] provides an answer how to avoid arbitrary distance selection.

In IB approach, given joint distribution $p(X, Y)$, the idea is search for a compact presentation of X, which preserves as much as possible information about the relevant variable (Y). Roughly speaking we are looking for document clusters (T) for document collection (X) so that clusters provide as much as possible information about the vocabulary (Y). The information between two variables A and B is measured with mutual information

$$I(A;B) = \sum_{a \in A} \sum_{b \in B} p(a)p(b|a) \log \frac{p(b|a)}{p(b}. \tag{1}$$

The IB principle means that, one wants to maximize the (mutual) information between document clusters and vocabulary (i.e. maximize $I(T;Y)$) and same time one wants to found as compact presentation of documents as possible (i.e. minimize $I(T;X)$). With these notation the score function F is

$$F = I(T;Y) - \beta I(T;X), \tag{2}$$

where β stands for Lagrange multiplier which eventually determines the tradeof between the quality ($I(T;Y)$) and the compactness ($I(T;X)$) of the presentation. The formal solution which maximize the score function F is given by equations:

$$\begin{aligned}
p(t|x) &= \frac{p(t)}{Z(\beta,x)} exp(-\beta D_{KL}(p(y|x)||p(y|t)), && (3)\\
p(y|t) &= \frac{1}{p(t)} \sum_{x \in X} p(t|x)p(x)p(y|x), && (4)\\
p(t) &= \sum_{x \in X} p(t|x)p(x), && (5)
\end{aligned}$$

where $Z(\beta,x)$ is a normalization factor and D_{KL} is a Kullbac-Leibler divergence.

In the case of the sequential clustering a prior assumption is that there is exactly K clusters in document collection X, and the score function is simply $F = I(T;Y)$. In sequential clustering one makes first an initial partition of documents to K clusters. The initial partition is done randomly assigning each document to one of the clusters. Step by step every document is drawn out of its current cluster and presented as a singleton cluster, which is then merged to cluster which minimize the distance between document and cluster centroid. It should be noted that in the case of *IB* method the choice of the distance was not arbitrary. The correct distance measure is found by solving the *IB* principle. By solving the score function of sequential IB method, we found that distance (distortion) between document and the centroid of document cluster is defined:

$$d(x,t) = (p(x) + p(t)) JS(p(y|x), p(y|t)), \tag{6}$$

where JS is a *Jensen-Shannon* divergence. This means that current document is assign to cluster, which minimize the distance measure and maximize the score function. The procedure is repeated over and over until the solution is converged. The score function has upper bound $I(X;Y)$ and because the value of the score function can either remain same or increase in each step then the convergence for (local) maximum is guaranteed. The algorithm can get trapped to local maxima and thus to make algorithm more robust the procedure is repeated n times with different random initializations. The solutio with highest score function is selected. The sequential IB method's time complexity has an upper bound $O(nLK|X|^2|Y|)^1$. The agglomerative IB method's, denoted as AIB, time complexity has an upper bound $O(|X|^3|Y|)$. Thus the proposed sequential method is more feasible for large document collections than the agglomerative IB method.

The performance of the *sIB* algorithm was compared to other clustering methods. The sIB was compared to AIB method, sequential Kullback-Leibler divergence based method, sequential L1-norm based method, sequential K-means method, and standard K-means method. The comparison and the evaluation were done against publicly available data sets. Experiments showed that the proposed sIB outperformed all reference unsupervised sequential methods. When compared to agglomerative IB method it was found that sequential method performed better than the agglomerative method and as stated above it is more feasible for large document collections.

[1]L is maximum number of iterations (user defined stopping criteria)