DISCRIMINATIVE CLUSTERING OF TEXT DOCUMENTS

Jaakko Peltonen, Janne Sinkkonen, and Samuel Kaski

Helsinki University of Technology
Neural Networks Research Centre
P.O. Box 9800, FIN-02015 HUT, Finland

ABSTRACT

Vector-space and distributional methods for text document clustering are discussed. Discriminative clustering, a recently proposed method, uses external data to find task-relevant characteristics of the documents, yet the clustering is defined even with no external data. We introduce a distributional version of discriminative clustering that represents text documents as probability distributions. The methods are tested in the task of clustering scientific document abstracts, and the ability of the methods to predict an independent topical classification of the abstracts is compared. The discriminative methods found topically more meaningful clusters than the vector space and distributional clustering models.

1. INTRODUCTION

Clustering texts to a smaller number of homogeneous groups is useful in mining, exploration, and summarization of text document collections, as well as in preprocessing for information retrieval.

Word order is often disregarded for computational reasons, and texts are considered "bags of words," finite-length multinomial samples. Topical content of the documents is then identified with the (underlying) multinomial distributions.

As the goal of clustering is to find homogeneous data subsets, how homogeneity is measured is crucial. For texts we should measure differences relevant for the topical content. A traditional solution has been to compile stop lists of irrelevant words, and to weight remaining words by estimated importance. The question has also been addressed by Latent Semantic Indexing (LSI) [1]. A probabilistic version of LSI is included in the comparisons of this paper.

In vector spaces choosing a measure of homogeneity is equivalent to choosing the feature selection and the distance measure, i.e. the metric. A recent method allows clusters to be constructed in terms of the primary data while the cluster homogeneity is still measured from other data within the clusters [8]. If suitable task-relevant auxiliary data is available, it can be used to indirectly define the homogeneity and to a degree circumvent the feature selection problem.

We apply this discriminative clustering method to scientific texts. The auxiliary data will be keywords from document authors. Assuming they have been chosen well, they signify what is relevant in the full text.

The method was introduced for vectorial data. We extend it to distributions, arguably more accurate representations for textual documents. Results are experimentally compared to standard vector-space and distributional probabilistic clustering methods. We test the ability of the various methods to discover topically homogeneous clusters, i.e., to predict a known, independent topical classification of the documents.

2. DOCUMENT CLUSTERING

There exist numerous clustering algorithms; here we focus on partitional clustering that divides the data space into a given number of partitions. Each text may be assigned to only one cluster, or more generally a membership function $y_j(\mathbf{n})$ may give the degree to which a document $\mathbf{n}$ belongs to the cluster j. Membership functions satisfy $\sum_j y_j(\mathbf{n}) = 1$ and $y_j(\mathbf{n}) \geq 0$.

Below we review some widely applied partitional text clustering methods and promising newer ones, used as references for the discriminative clustering methods in Section 3.

2.1. Mixture Model in a Vector Space

Salton [7] introduced the vector space model (VSM) to the information retrieval field. Text documents are represented as points $\mathbf{x}$ in a vector space. Each word corresponds to a dimension of the space; the coordinate of a document along the dimension is determined by the number of occurrences of the word in the document. Document similarity is measured by the angle or inner product of the document vectors.

The mixture density model [5] is applicable to the VSM. Assume a document is produced by one of many generators

(soft clusters). The data density is modeled as their mixture,

$$p(\mathbf{x}) = \sum_j p(\mathbf{x}|\mathbf{m}_j)p_j \ , \qquad (1)$$

where p_j is the probability of cluster j. The parameters $\mathbf{m}_j$ and p_j are optimized by maximizing the model likelihood, and cluster memberships $p(\mathbf{m}_j|\mathbf{x})$ can be computed by the Bayes rule.

In VSM the normalized documents lie on a hypersphere, so the appropriate density estimator is a mixture of von Mises-Fisher kernels, the hypersphere analogs of Gaussians: $p(\mathbf{x}|\boldsymbol{\theta}_j) = (Z(\kappa))^{-1}\exp(\kappa\mathbf{x}^T\mathbf{m}_j)$, where κ governs the spread of the kernel and Z normalizes p to a proper density. We do not apply term weighting, so $\mathbf{x}$ is simply $\mathbf{n}$ normalized to unit length.

2.2. Distributional Clustering of Co-occurrence Data by the Information Bottleneck

The Information Bottleneck method [6, 9] can be used for clustering documents $\mathbf{n}_l$. The documents are first converted into a distributional form by $q_{lk} = n_{lk}/\sum_r n_{lr}$. A (soft) partitioning of the documents to a set of clusters j is then sought by minimizing a cost function, motivated by information theory but expressable as

$$E_{DIS} = \sum_l \sum_j p_j(\mathbf{q}_l) D_{KL}(\mathbf{q}_l, \boldsymbol{\theta}_j) - \beta^{-1} I \, , \qquad (2)$$

where $p_j(\mathbf{q}_l)$ represent the cluster memberships (in the form of probabilities, summing to unity over j), D_{KL} is the Kullback-Leibler divergence between a document $\mathbf{q}_l$ and a prototype $\boldsymbol{\theta}_j$, and I denotes the mutual information between the generated document clusters, regarded as a random variable, and the documents themselves. Variational optimization leads to clusters of the general form quite similar to (7), leaving the prototypes $\boldsymbol{\theta}_j$ and the (prior) probabilities of the clusters to be fitted to the data. The parameter β chooses a compromise between cluster smoothness and minimization of the average distortion.

In this paper we did not implement (2) but the related method presented below.

2.3. Mixture Models for Co-occurrence Data

A generative probabilistic model called the Asymmetric Clustering Model (ACM; [3]) is closely related to the distributional clustering method above. It has been shown [3] that obtaining the maximum likelihood solution of ACM is equivalent to minimizing

$$E_{ACM} = \sum_l n(\mathbf{q}_l) \sum_j p_j(\mathbf{q}_l) D_{KL}(\mathbf{q}_l, \boldsymbol{\theta}_j) \ , \qquad (3)$$

where $n(\mathbf{q}_l)$ is the empirical frequency of document $\mathbf{q}_l$.

In ACM each document is probabilistically assigned to the clusters (cluster memberships are *a priori* unknown). An alternative is to directly model co-occurrence patterns of words and documents. In the Separable Mixture Model (SMM; [3]), also called probabilistic LSI, the co-occurrence probability of word w_k and document $\mathbf{q}_l$ is modeled by

$$p(\mathbf{q}_l, w_k) = \sum_j p_j p(\mathbf{q}_l|u_j)p(w_k|u_j) \ , \qquad (4)$$

where u_j denotes the cluster j, and p_j is the probability of cluster j. All probabilities here are parameters, optimized by maximizing the likelihood with the EM algorithm.

SMM is not designed to be a clustering method; it decomposes co-occurrences probabilistically into factors. It can be used for clustering by regarding the factors as clusters. The cluster probabilities for a document can be computed by the Bayes rule.

3. DISCRIMINATIVE CLUSTERING

A recent clustering principle aims to implicitly find an optimal way to measure data similarity [4, 8]. We call this principle discriminative clustering since it incorporates discriminative elements into a clustering task.

In general, clustering aims to maximize within-cluster similarity or homogeneity. Discriminative clustering is applicable when primary samples can be paired with discrete auxiliary labels.

The auxiliary data is supposed to be a canonical indicator of important variation in the primary data. Inhomogeneities in the primary data are noted only if they are associated to variation in the conditional auxiliary distributions.

The homogeneity measure is within-cluster similarity of the auxiliary data distributions. However, the clusters are defined in terms of the primary data. The auxiliary data only guides the optimization. Given a clustering, new samples can be clustered without any auxiliary data.

Previously, a vector-space clustering algorithm has been presented [8] which we denote Vector-space Discriminative Clustering (VDC). This general-purpose clustering method, applied to texts in vector form, works well in practice. Simplifying assumptions are made, though, and taking into account the distributional nature of the texts could improve results. Here we derive a distributional version, arguably more compatible with text documents and the "bag of words."

3.1. Discriminative Clustering of Texts

Assume that the documents are generated by multinomial distributions with parameter vectors $\mathbf{q}$, and denote the auxiliary samples by c. We construct a parameterized partitioning $y_j(\mathbf{q})$ into the distribution space. The partitioning is

softened due to computational reasons, by allowing several nonzero memberships $y_j(\mathbf{q})$ for $\mathbf{q}$.

Discriminative clustering generalizes vector quantization (VQ) that represents data by prototypes and minimizes the caused distortion. We measure the distortion between the conditional distributions $p(c|\mathbf{q})$ and distributional prototypes $\boldsymbol{\psi}_j$. The average distortion then is

$$E' = \sum_j \int y_j(\mathbf{q};\boldsymbol{\Theta}) D_{KL}(p(c|\mathbf{q}),\boldsymbol{\psi}_j) p(\mathbf{q}) d\mathbf{q} \; . \qquad (5)$$

Here $p(\mathbf{q})$ is the (unknown) sampling distribution of our data. The distributional prototypes are re-parameterized by

$$\psi_{ji} = \frac{\exp \gamma_{ji}}{\sum_m \exp(\gamma_{jm})} \qquad (6)$$

to keep them summed up to unity. The membership functions y_j in the distributional space are parameterized as (normalized) Gaussians, with the distance measured by the Kullback-Leibler divergence D_{KL}, a natural choice for the now distributional document space. This yields

$$y_j(\mathbf{q};\boldsymbol{\Theta}) = e^{-\kappa D_{KL}(\mathbf{q},\boldsymbol{\theta}_j))} / Z(\boldsymbol{\Theta}) \; , \qquad (7)$$

where $Z(\boldsymbol{\Theta})$ ensures $\sum_j y_j(\mathbf{q}) = 1$, and $\boldsymbol{\Theta}$ denotes the collection of all parameters of the membership functions.

It is easy to see resemblance to the information bottleneck method (Section 2.2): e.g. the cluster memberships y_j have a very similar functional shape [9]. One view to the differences[1] is that distributional clustering performs constrained (soft) "vector quantization" in the space of multinomial distributions, while discriminative distributional clustering finds partitions relevant for the auxiliary variable C.

3.2. Noise Model

In practice $\mathbf{q}$ are unknown, and documents $\mathbf{n}$ have finite length. Here we postulate a noise model that takes this fact into account, and propose a tractable approximation essentially equivalent to (5).

In the "bag of words" assumption, $\mathbf{n} \sim p(\mathbf{n}|\mathbf{q}, M)$ with parameters $\mathbf{q}$ of the multinomial model M. The posterior parameter distribution is $p(\mathbf{q}|\mathbf{n}, M) \propto p(\mathbf{n}|\mathbf{q}, M)p(\mathbf{q}|M)$, where $p(\mathbf{q}|M)$ is a prior.

To take uncertainty about $\mathbf{q}$ into account, instead of (5) we could minimize the distortion averaged over parameters:

$$E'' = \iint \sum_j y_j(\mathbf{q};\boldsymbol{\Theta}) d_j(\mathbf{q}) \, p(\mathbf{q}|\mathbf{n}, M) d\mathbf{q} \, p(\mathbf{n}) d\mathbf{n} \; , \quad (8)$$

where $d_j(\mathbf{q}) = D_{KL}(p(c|\mathbf{q}), \boldsymbol{\psi}_j)$.

[1] The distributional clustering model could be applied to documents and keywords as well, but then the results would not be readily applicable to new full-text documents without keywords.

Let us approximate (8). With a (conjugate) Dirichlet prior for the $\mathbf{q}$ symmetric across the words, the posterior $p(\mathbf{q}|\mathbf{n}, M)$ becomes proportional to $\Pi_k q_k^{n_k}$, with its mode at $\hat{q}_k(\mathbf{n}) \equiv n_k / \sum_r n_r$. If the posterior is approximated by its mode in the average distortion (8), the distortion simplifies into

$$E_{DDC} = \int \sum_j y_j(\hat{\mathbf{q}}(\mathbf{n});\boldsymbol{\Theta}) D_{KL}(p(c|\mathbf{n}), \boldsymbol{\psi}_j) p(\mathbf{n}) d\mathbf{n} \; . \qquad (9)$$

This is equal to (5), with the word frequencies of documents normalized to approximate distributions. We call this model Discriminative Distributional Clustering (DDC).

3.3. The DDC Algorithm

The partitioning is optimized by minimizing the cost function (9) with respect to both prototype sets, $\boldsymbol{\theta}_j$ and $\boldsymbol{\psi}_j$.

It can be shown that this can be done by a stochastic approximation algorithm that iterates the following steps:

1. At iteration t, sample a labeled text document $(\mathbf{n}(t), c_i)$ from the distribution $p(\mathbf{n}, c)$ (in practice: randomly from the data). Below i denotes the index of the the auxiliary value. Denote $\hat{\mathbf{q}}_t = \hat{\mathbf{q}}(\mathbf{n}(t))$.
2. Sample clusters j and l from the distribution $\{y_k(\hat{\mathbf{q}}_t)\}_k$.
3. Adapt the parameters according to

$$\boldsymbol{\beta}_l(t+1) = \boldsymbol{\beta}_l(t) - \alpha(t)(\boldsymbol{\theta}_l(t) - \hat{\mathbf{q}}_t) \log \frac{\psi_{ji}(t)}{\psi_{li}(t)} \qquad (10)$$

$$\gamma_{lm}(t+1) = \gamma_{lm}(t) - \alpha(t)[\psi_{lm}(t) - \delta_{mi}] \qquad (11)$$

The $\boldsymbol{\theta}_j$ have been re-parameterized by setting $\theta_{ji} = \exp(\beta_{ji}) / \sum_k \exp(\beta_{jk})$, and $\boldsymbol{\psi}_j$ by (6). Due to symmetry, it is possible (and apparently advantageous) to adapt the parameters twice for each t by swapping j and l in (10) and (11) for the second adaptation. The positive and gradually decreasing learning coefficient $\alpha(t)$ should in principle fulfill the conditions of the stochastic approximation theory: $\sum_t \alpha(t) = \infty$ and $\sum_t \alpha^2(t) < \infty$.

4. EMPIRICAL COMPARISON

In this section we compare the models empirically: the vector space mixture model (called vMF-M below), the generative co-occurrence models ACM and SMM, and the discriminative clustering models in the vector (VDC) and distribution space (DDC).

We hypothesize that discriminative clustering models discover more essential structure in the data and outperform the other methods in topical clustering. This is measured by the ability of the clusterings to predict independent topical categories produced by informaticians.

It is also of note whether distributional models outperform the more heuristic vector space models; here interesting pairs are ACM/SMM vs. vMF-M, and DDC vs. VDC.

4.1. The Data and Feature Selection

The data were scientific abstracts from the INSPEC database. Documents were collected from nine partially overlapping INSPEC topic categories. Topic categories were only used in the final phase to compare the methods.

All algorithms clustered textual documents consisting of the free text and the title fields of the abstracts. Words were converted to base form, and occurrences were counted. The discriminative methods used the keywords field of the abstracts as the auxiliary data; keywords are descriptive words for the documents given by the original authors. All keywords in the set of 1500 most frequent ones were accepted. For VDC and DDC, cases of multiple keywords per document were assimilated by minimizing the average cost over the keywords.

We ran two sets of experiments with different preprocessing. The first ("random features") used no prior information about word relevance. 500 words were picked randomly from words with over 50 occurrences in the corpus.

The second experiment ("IDF-picked features") used more prior information. Words in a stop-list of 1335 words were discarded, after which 500 words with largest IDF weights were chosen. IDF is the inverse of how many documents the word occurs in.

The final data contained all documents with at least one of the 500 words, yielding 53,613 documents for "random features," and 13,162 documents for "IDF-picked features."

4.2. Optimization of the Methods

The number of clusters was set to nine for all models. The discriminative models were trained by 10^6 on-line iterations of stochastic approximation, during which $\alpha(t)$ was decreased piecewise-linearly to zero. The γ were updated with a higher $\alpha(t)$ than the other parameters. The precise values were chosen based on preliminary experiments.

ACM and SMM were trained to convergence by the EM algorithm, and then by deterministic annealing iterations until convergence, as recommended in [2]. The vMF-M was optimized to convergence by the EM algorithm.

The dispersion parameter κ of vMF-M, DDC, and VDC, and the annealing parameters of ACM and SMM, were chosen by validation: the models were optimized for a validation set equal in size to the training set.

To keep the optimization of the annealing parameter comparable to the cross-validation of parameters in the other models, instead of varying it within a run as suggested in [2] it was kept constant and the EM algorithm was run until convergence.

Data set	Model	Mean	STD
Random features	DDC	**0.56**	0.023
	VDC	0.47	0.022
	vMF-M	0.26	0.014
	ACM	0.48	0.015
	SMM	0.12	0.006
IDF-picked features	DDC	0.58	0.060
	VDC	**0.80**	0.048
	vMF-M	0.18	0.023
	ACM	0.23	0.023
	SMM	0.08	0.015

Table 1: Results by 10-fold cross-validation for all models on the two sets. Mean and standard deviation (STD) of empirical mutual information (in bits) between clusters and categories. The best results are shown in bold. The models shown are Discriminative Distributional Clustering (DDC), Vector-space Discriminative Clustering (VDC), vector-space mixture model (vMF-M), the Asymmetric Clustering Model (ACM), and the Separable Mixture Model (SMM).

4.3. Evaluation

The models were compared by how well they were able to extract independent topically meaningful clusters. The criterion was their ability to predict the nine INSPEC categories of the abstracts. Category information was not used in training.

The performance of the models was measured as the empirical mutual information between the extracted clusters and the topic categories. The (empirical) mutual information was estimated from test data not used in training.

Empirical mutual information is positively biased for small samples. We reduced the bias by measuring a "soft" mutual information I. The conditional probabilities $\hat{p}(v_j|\mathbf{n}_l)$ of clusters v_j given the document l were used instead of assigning the document to a single cluster. Technically, we computed

$$I = \sum_i \sum_j f_{ij} \log \frac{f_{ij}}{\sum_l f_{il} \sum_k f_{kj}} ,$$

where the experimental relative frequencies f_{ij} are

$$f_{ij} = \frac{\sum_{k:c(\mathbf{n}_k)=c_i} \hat{p}(v_j|\mathbf{n}_k)}{\sum_{i,j} \sum_{k:c(\mathbf{n}_k)=c_i} \hat{p}(v_j|\mathbf{n}_k)} .$$

Here c_i denotes the ith topic category, $c(\mathbf{n}_k)$ is the topic category of sample $\mathbf{n}_k$, and j indexes the clusters.

4.4. Results and Demonstration

The performance of the models for the two data sets is shown in Table 1. The discriminative methods (DDC and VDC)

Cluster	Titles of sample documents
151	"A genetic algorithm approach to Chinese handwriting normalization" "Size normalization in on-line unconstrained handwriting recognition"
278	"On the security of the McEliece public key cryptosystem" "An implementation of an elliptic curve cryptosystem"
243	"Movement and memory function in biological neural networks" "An analog retina for edge detection"

Table 2: Sample documents mapped to three clusters found by the VDC algorithm using "IDF-picked features."

outperform unsupervised models: DDC attains the highest mutual information for the "random features" and VDC for the "IDF-picked features." The differences to the next best models are significant (McNemar test, $p<0.005$).

ACM consistently outperforms the unsupervised vMF-M ($p<0.005$) and is surprisingly good for "random features", where its results are roughly equal with the supervised VDC model (not significantly different) but below DDC. Surprisingly, SMM was the worst for both data sets ($p<0.005$).

The comparison between the discriminative algorithms is interesting: VDC was better for IDF-picked features, while the distributional version (DDC) worked better for randomly chosen features. More detailed investigation is needed; one possible reason is data sparseness in the former case.

To demonstrate the discriminative algorithms we computed 400 clusters for a subset of about 9000 documents of the "IDF-chosen features" set. The dispersion parameter κ of the VDC model was selected by using the remaining, about 4000 documents, as a validation set.

Sample document titles from three clusters are shown in Table 2. Cluster 151 has articles on handwriting recognition, 278 on cryptosystems, and 243 on biological and artificial neural networks. All articles in the clusters were not as homogeneous, though. Typically there seemed to be articles from about 1–3 topics in each cluster.

5. DISCUSSION

We have shown that discriminative clustering improves text clustering results. The clusters are more closely related to relevant categories given by human experts, even though the categories were not used in training.

The full-text clustering was guided by auxiliary data, here keywords from document authors. The clusters become homogeneous in the keywords and discriminate them well. Still, the primary, full-text space is clustered, and clusters are defined even for documents without keywords.

An alternative would be to estimate the joint density of documents, words, and classes (keywords), and define the clusters by marginalizing the estimated density. For fixed resources, however, the result would probably be suboptimal for other purposes besides joint density estimation.

Here we sought a fixed number of clusters. Criteria for choosing the number of clusters need be developed later. An alternative would be to build a large cluster set and summarize it by e.g. agglomeration.

6. ACKNOWLEDGMENTS

This work was supported in part by the Academy of Finland, grants 52123 and 1164349, and a contract with Reach-U Oyj.

7. REFERENCES

[1] S. Deerwester, S. T. Dumais, G. W. Furnas, and T. K. Landauer, Indexing by latent semantic analysis, *Journal of the American Society for Information Science* 41, pp. 391-407, 1990.

[2] T. Hofmann, Probabilistic latent semantic analysis, in *Proceedings of the Fifteenth Conference on Uncertainty in Articial Intelligence*, Morgan Kaufmann Publishers, San Francisco, CA, pp. 289-296, 1999.

[3] T. Hofmann and J. Puzicha, Statistical models for co-occurrence data, A.I. Memo 1625, MIT, 1998.

[4] S. Kaski and J. Sinkkonen, Metrics that learn relevance, in *Proceedings of the International Joint Conference on Neural Networks*, volume V, IEEE Service Center, Piscataway, NJ, pp. 547-552, 2000.

[5] G. J. McLachlan and K. E. Basford, *Mixture Models. Inference and Applications to Clustering*, Marcel Dekker, New York, NY, 1988.

[6] F. Pereira, N. Tishby, and L. Lee, Distributional clustering of English words, in *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, ACL, Columbus, OH, pp. 183-190, 1993.

[7] G. Salton and M. J. McGill, *Introduction to modern information retrieval*, McGraw-Hill, New York, 1983.

[8] J. Sinkkonen and S. Kaski, Clustering based on conditional distributions in an auxiliary space, *Neural Computation* 14, pp. 217-239, 2002.

[9] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, in *The 37th Annual Allerton Conference on Communication, Control, and Computing*, Urbana, Illinois, 1999.