Robust Initialization for Learning Latent Dirichlet Allocation

Pietro Lovato[1(✉)], Manuele Bicego[1], Vittorio Murino[2], and Alessandro Perina[2]

[1] Department of Computer Science, University of Verona,
Strada le Grazie 15, 37134 Verona, Italy
pietro.lovato@univr.it

[2] Pattern Analysis and Computer Vision (PAVIS),
Istituto Italiano di Tecnologia (IIT), Via Morego 30, 16163 Genova, Italy

Abstract. Latent Dirichlet Allocation (LDA) represents perhaps the most famous topic model, employed in many different contexts in Computer Science. The wide success of LDA is due to the effectiveness of this model in dealing with large datasets, the competitive performances obtained on several tasks (e.g. classification, clustering), and the interpretability of the solution provided. Learning the LDA from training data usually requires to employ iterative optimization techniques such as the Expectation-Maximization, for which the choice of a good initialization is of crucial importance to reach an optimal solution. However, even if some clever solutions have been proposed, in practical applications this issue is typically disregarded, and the usual solution is to resort to random initialization.

In this paper we address the problem of initializing the LDA model with two novel strategies: the key idea is to perform a repeated learning by employ a topic splitting/pruning strategy, such that each learning phase is initialized with an informative situation derived from the previous phase.

The performances of the proposed splitting and pruning strategies have been assessed from a twofold perspective: *i)* the log-likelihood of the learned model (both on the training set and on a held-out set); *ii)* the coherence of the learned topics. The evaluation has been carried out on five different datasets, taken from and heterogeneous contexts in the literature, showing promising results.

Keywords: Topic models · LDA · Split · Prune · Expectation-Maximization

1 Introduction

Topic models represent an important and flexible class of probabilistic tools, originally introduced in the Natural Language Processing community [5,6,20]. Their main goal is to describe text documents, based on word counts, abstracting the *topics* the various documents are speaking about. Recently, the importance

A. Feragen et al. (Eds.): SIMBAD 2015, LNCS 9370, pp. 117–132, 2015.
DOI: 10.1007/978-3-319-24261-3_10

of topic models has drastically grown beyond text, and they have been exported as a versatile tool to model and solve a huge variety of tasks in different contexts [1,8,21,25,30,34,36]. Their wide usage is motivated by the competitive performances obtained in very different fields, by their expressiveness and efficiency, and by the interpretability of the solution provided [9]. Among others, Latent Dirichlet Allocation (LDA) [6] is the most cited and famous topic model. The key idea of LDA is that a document may be characterized by the presence of a small number of topics (e.g. sports, finance, politics), each one inducing the presence of some particular words that are likely to co-occur in the document; the total number of topics expected to be found in the *corpus* of documents is a fixed quantity decided beforehand. From a probabilistic point of view, a topic is a probability distribution over a fixed dictionary of words: for example, a topic about sports would involve words like "match" or "player" with high probability.

The parameters of the model are *learned* from a set of training objects: however, the learning problem is intractable [6], and is therefore tackled using approximate optimization techniques such as the variational Expectation-Maximization (EM [11,17]). The EM is an iterative algorithm that, starting from some initial values assigned to the afore-described probabilities, maximizes the log likelihood of the model until convergence is reached. The choice of such initial values is a critical issue because the EM converges to a local maximum of the log likelihood function [35], and the final estimate depends on the initialization.

From a very general point of view, different robust variants of the EM algorithm have been proposed in the past ([13,33], just to cite a few); nevertheless, in most practical applications where the LDA model is employed, this initialization problem is overlooked, with most solutions starting the EM iterations from a random solution; this is usually motivated by the already appropriate performances of the method. Only few papers explicitly addressed the EM initialization issue in the LDA case: the authors of [15] proposed to employ a clustering step using the k-means as initial solution for the EM; in [14], a method based on the SVD decomposition is proposed. These methods have been originally proposed for a slightly different topic model called PLSA [20], but can be easily adapted for LDA. More often, workarounds are employed at experimental level: in some cases, the learning is repeated several times, and average performances are reported [19]. In other cases, the learning is repeated several times, and the model with the highest log likelihood is retained [3] (also employed in other EM-based techniques, such as Gaussian mixtures clustering [26]).

In this paper we contribute to this context, by proposing two novel strategies for the initialization of the LDA training that specifically exploit the intrinsic characteristics of the LDA model and the information therein. Both approaches share the same structure: start by learning a model with an extremely small (or an extremely large) number of topics, proceeding with consecutive operations of splitting (pruning) of the topics, until the desired number of topics is reached; each learning phase is initialized with an informative situation derived from the previous phase. The pruning strategy takes inspiration from the observation that, when the number of topics is extremely large, the dependency from the initialization of the final estimate is much weaker than when the number of topics is close to the optimum [4,16,28]. On the other hand, the splitting approach exploits reasoning derived for divisive clustering algorithms, where it has been shown that such a strategy is useful when the size of the dataset is particularly high [7,12,31]. In both cases, the approach initializes these "extreme" models at random, and use the learned estimates to initialize a new model with a number of topics closer to the desired one. To choose which are the best topics to split/prune, we exploit a quantity which can be readily extracted from the learned LDA: the prior Dirichlet probability, which can be thought of a number indicating the "importance" of each individual topic. This quantity is intrinsic in the LDA formulation, and is not exploited by the methods described in [14,15].

The proposed splitting and pruning strategies have been extensively tested on 5 datasets, taken from heterogeneous applicative contexts where LDA has already been successfully employed. Benefits and merits of both techniques are discussed, as well as the situations where one seems better suited over the other. Experimental results confirm the usefulness of initializing the LDA model with the proposed approach (*i*) in terms of the model log likelihood (evaluated both on the training set and on a held out, testing set) and (*ii*) in terms of the coherence and the interpretability of the learned topics.

The remainder of the paper is organized as follows: Sect. 2 gives some background notions on the LDA model, whereas Sect. 3 details the proposed strategies of robust initialization. Sect. 4 contains the experimental evaluation, and the discussion of the obtained results. Finally, in Sect. 5 conclusions are drawn and future perspectives envisaged.

2 Background: Latent Dirichlet Allocation

In the general LDA formulation, the input is a set of D objects (e.g. documents), represented as "bag of words" vectors $\mathbf{c}^d$ [27]. The bag of words is a representation particularly suited when the object is characterized (or assumed to be characterized) by the repetition of basic, "constituting" elements w, called words. By assuming that all possible words are stored in a dictionary of size N, the bag of words vector $\mathbf{c}^d$ for one particular object (indexed by d) is obtained by counting the number of times each element w_n of the dictionary occurs in d.

In LDA, the presence of a word w_n in the object d is mediated by a latent *topic* variable, $z \in Z = \{z_1,...,z_K\}$. The joint probability of the model variables is:

$$p(w_n, z_k, \theta^d) = p(\theta^d|\alpha)p(z_k|\theta^d)p(w_n|z_k, \beta) \tag{1}$$

In other words, the topic z_k is a probabilistic co-occurrence of words encoded by the distribution $p(w_n|z_k, \beta)$, $w = \{w_1,..., w_N\}$, where β represents, in tabular form, the probability of word w_n being "involved" in topic z_k. The variable $\theta_k^d = p(z_k|\theta^d)$ represents the proportion of topics in the object indexed by d; finally $p(\theta|\alpha)$ is a Dirichlet prior indicating the probability of selecting a particular mixture of topics: α_k can be seen as a measure of the prior "importance" of each topic. From this, the process that generates an object is defined as follows.

Table 1. Summary of the LDA distributions.

Name	Distribution	Parameter	Dimensionality
$p(\theta^d\|\alpha)$	Dirichlet	α	$1 \times K$
$p(z_k\|\theta^d)$	Multinomial	θ	$K \times D$
$p(w_n\|z_k,\beta)$	Multinomial	β	$N \times K$

First, the proportion of topics θ that will compose the object is generated from the Dirichlet $p(\theta|\alpha)$; then, a topic z_k is drawn from the distribution $p(z|\theta)$, and from this topic a word is selected according to the probabilities in β. Finally, the process is repeated, by selecting another topic and another word, until the whole object is generated. A summary of the distributions involved in the LDA formulation, as well as their parameter dimensionality, is summarized in Table 1.

Learning the LDA model requires to estimate the parameters α and β from a set of training data, in order to maximize the likelihood $\mathcal{L}$, defined as

$$\mathcal{L} = p(D|\alpha,\beta) = \prod_{d=1}^{D} \int_{\theta^d} p(\theta^d|\alpha) \left(\sum_{k=1}^{K} \prod_{n=1}^{N} \Big(p(z_k|\theta^d) p(w_n|z_k,\beta) \Big)^{c_n^d} \right) \tag{2}$$

Since this function is intractable [6], such parameters are learned using a variational Expectation-Maximization algorithm (EM) [17]. The EM iteratively learns the model by minimizing a bound $\mathcal{F}$ (called the *free energy* [17]) on the negative log likelihood, by alternating the E and M-step. A detailed derivation of the EM algorithm for LDA is out of the scopes of this paper (interested readers can refer to the original LDA paper [6]): intuitively, the derivation yields the following iterative algorithm:

1. Initialize α and β
2. **E-step:** for each object in the training set, estimate the posterior probability $p(\theta, \mathbf{z} \mid \mathbf{c}^d, \alpha, \beta)$ (obtained by using Bayes' law from the likelihood formula in Eq. 2). Unfortunately, obtaining such estimate proved to be intractable [6], and so an approximate form of the posterior is estimated.
3. **M-step:** minimize the free energy bound with respect to the model parameters α and β. This corresponds to find a maximum likelihood estimate for each object, under the approximate posterior which is computed in the E-step.
4. Repeat the steps 2 and 3 until some convergence criterion (usually, a small variation in the free energy between two consecutive iterations) is met.

Summarizing, the EM is an iterative algorithm that, starting from some initial values assigned to the parameters α and β, refines their estimates by maximizing the log likelihood of the model until convergence is reached. As outlined in the introduction, the choice of such initial values is a critical issue because the EM converges to a local maximum of the free energy function [35]: the final estimate depends on the initialization.

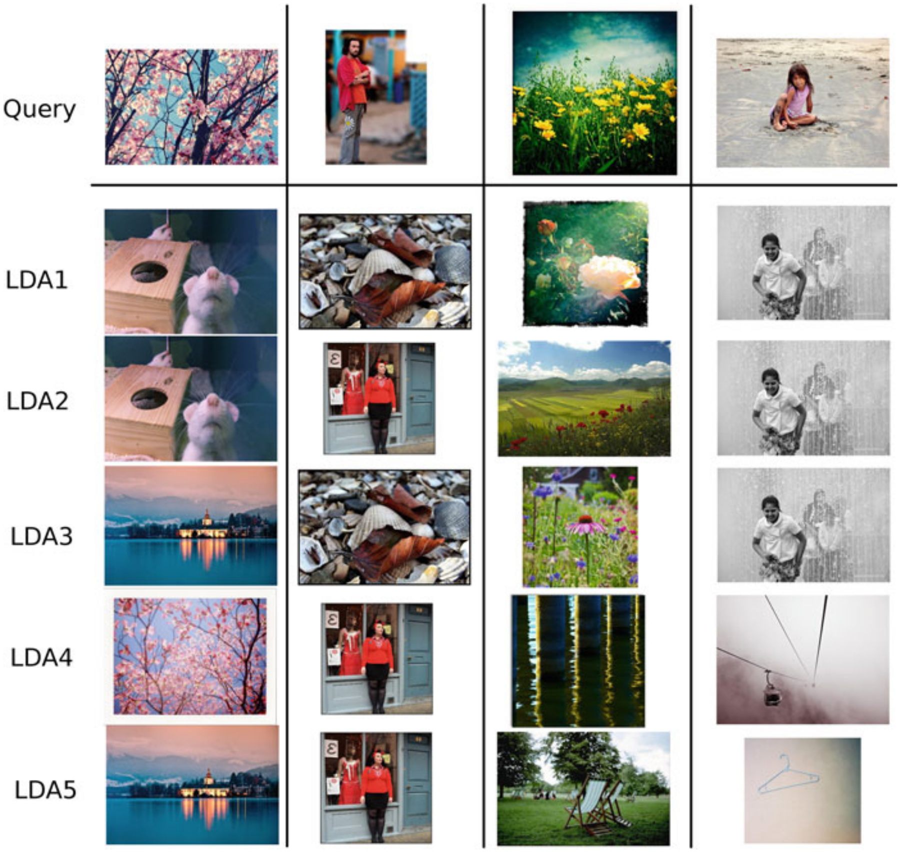

Fig. 1. The top-most row shows some query images we selected: 5 independent runs of the LDA model (initialized at random) produce very different retrieved images, presented under each query image.

Even if this problem is known, most practical systems initialize the EM iterations at random. This may lead to very poor results: let us clarify this point with a simple toy example, inspired by the framework of [8]. In that paper, the goal was to classify a query image into a scene category (e.g. mountain, forest, office): first, the LDA is learned on a training set, and each training image d is projected in the topic space through the vector θ^d. Then, the query image d_{test} is also projected in the topic space via an E-step, and its vector $\theta^{d_{\text{test}}}$ is estimated. The retrieval step can be carried out by simply showing the nearest neighbor, computed for example using the euclidean distance between $\theta^{d_{\text{test}}}$ and the training θ^d. In our simple example, we devised the same retrieval strategy on a recent dataset of images collected from Flickr[1]: in particular we learned 5 LDA models – in each case starting with a different random initialization – on a

[1] More details on the dataset, called PsychoFlickr, can be found in [10].

given set of roughly 10000 images. Then, given a query image, we retrieved the most similar by using the five different models; the expectation, if the LDA is well trained, is to extract in all the 5 cases the same image. In Fig. 1 we show some results: it can be immediately noted that, in different cases, the retrieved images are diverse, in some cases also visually rather unrelated to the query.

3 The Proposed Approach

As stated in the introduction, the goal of this paper is to derive two robust initialization techniques for the parameters α and β of LDA, by exploiting the intrinsic characteristics and the information derived from the model. In this section the two strategies, that we term *splitting* and *pruning*, will be detailed. Intuitively, the idea is to initialize at random the LDA model designed with an extremely small (for the splitting strategy), or an extremely large (for the pruning strategy) number of topics, performing a series of splitting or pruning operations until the chosen number of topics is reached.

In the following, the proposed initialization techniques are detailed.

3.1 LDA Initialization by Pruning

Suppose that the goal is to learn the LDA model with K topics. First, we propose to learn a model with an extremely large number of topics K_{large}, initialized at random: the idea behind this approach is that this first run of the EM, due to the excessive number of topics (at the extreme, equal to the number of training documents D), is less sensitive to initialization [4,16,28]. After the model is learned, we select a candidate topic to prune, update α and β, and repeat the learning starting with this new configuration. Of course, the crucial problem is to decide which topic to prune. To make this choice, we look at the

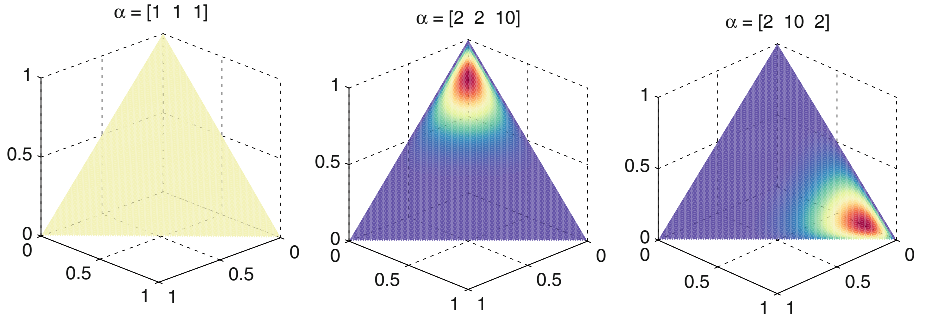

Fig. 2. Effects on the α parameter on the sampled topic proportions θ. The triangular region correspond to the simplex where the θ probability distributions live, with the edges of the triangle corresponding to the θ distribution where only one topic is present with probability 1. Note that high values of α for a particular topic k "move" the proportions θ to be concentrated towards k.

learned parameter α of the prior Dirichlet distribution. Intuitively, a high value of α_k indicates that a specific topic distribution θ – where k is highly present – is more likely to appear in the corpus of training objects. On the contrary, a low value of α_k indicates that k is overall scarcely present – Fig. 2 depicts a graphical illustration of this idea.

For the above mentioned reason, it seems reasonable to consider as the least interesting topic, i.e. the topic to prune, the topic $\hat{k}$ with the lowest corresponding α, i.e.

$$\hat{k} = \arg\min_{k} \alpha_k \tag{3}$$

In practice, pruning a topic $\hat{k}$ implies (i) to remove its $\alpha_{\hat{k}}$ value, and (ii) to remove the whole vector of probabilities from β, i.e. $\beta_{n,\hat{k}} = p(w_n|z_{\hat{k}})$ for each n. This is graphically pictured in the left part of Fig. 3. After the pruning, the remaining parameter vectors α and β can provide a good starting point for the learning of a new LDA, where the number of topics is decreased by one. This is reasonable because we are making simple modifications to a good solution (the model has already converged). Finally, the learning is repeated until K topics are obtained.

From a practical point of view, it is interesting to notice that it is not necessary to prune one topic at a time: the learned prior α can be used to rank topics, from the least to the most important, and an arbitrary number of unimportant topics can be pruned before repeating the learning procedure. The main advantage is that computational cost is reduced, because less LDA models have to be learned; however, this can deteriorate the quality of the final solution.

Finally, we can draw a parallelism between our approach and an agglomerative hierarchical-type clustering scheme: we start from a large number of topics and evolve by decreasing such number until the desired one is reached.

3.2 LDA Initialization by Splitting

Contrarily to the pruning approach, the idea behind the splitting strategy is to initialize at random an LDA model with an extremely small number of topics K_{small}, and proceeding by splitting one topic at a time into two new topics.

From a clustering perspective, the splitting approach can be seen as a divisive (or top-down) hierarchical-type scheme: starting from a small number of clusters, the process evolves towards a greater number of clusters. Divisive clustering algorithms proved to be particularly appropriate when the size of the data is particularly high [7,12,31], and seem therefore a promising strategy to investigate in this context. Once the first model with K_{small} topics is learned, we employed – as for the pruning strategy – the α prior in order to decide the topic to split. In particular, the idea is that a high value of α indicates an overall highly present topic in the training set. From the divisive clustering perspective, these topics are the "largest", clustering together many words and summarizing most of the objects. For this reason, we propose to split the topic $\hat{k}$ such that

$$\hat{k} = \arg\max_{k} \alpha_k \tag{4}$$

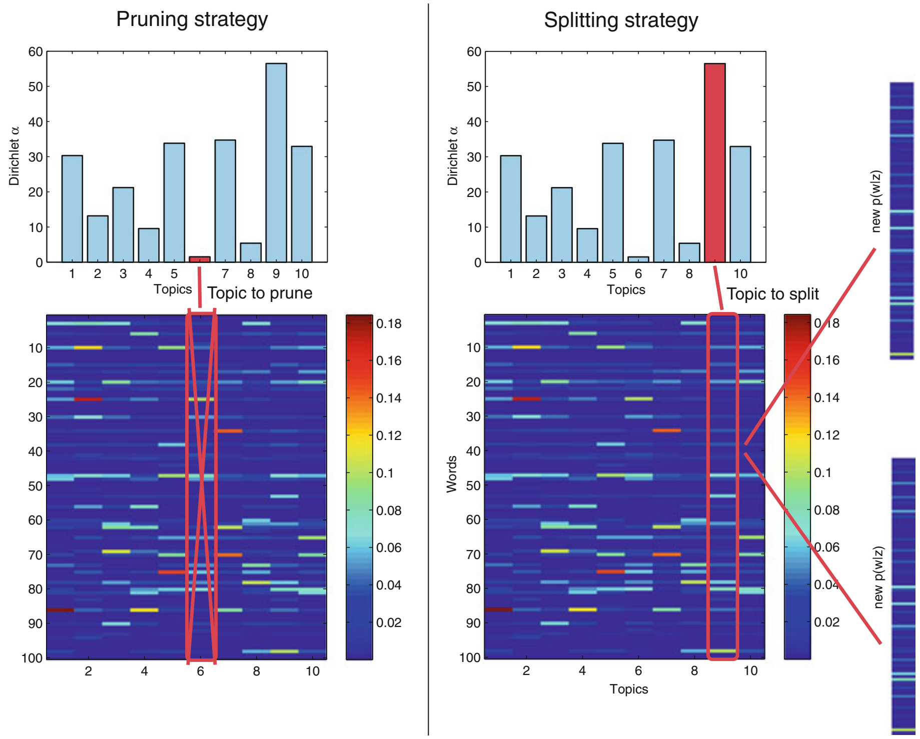

Fig. 3. (left) Summary of the pruning strategy. The top bar graph represents the learned α parameter after EM convergence. The candidate topic to prune is the one with the lowest value of α. On the bottom, the β probabilities are graphically depicted, with a brighter, red color indicating a higher probability of a particular word belonging to a particular topic (each column corresponds to a topic). This topic is pruned by simply removing the corresponding column from the β probabilities. (right) Summary of the splitting strategy. Given a learned LDA, the topic to split the one with the highest value of α. A small amount of Gaussian noise is applied to each copy of the splitted topic (Color figure online).

In practice, splitting a topic $\hat{k}$ implies to substitute the topic $\hat{k}$ with two topics $\hat{k}_1$ and $\hat{k}_2$ such that: (i) the β probability of $\hat{k}_1$ and $\hat{k}_2$ are equal to the β of $\hat{k}$ plus a small amount of Gaussian noise (a simple normalization is also applied so that such probabilities add up to 1); (ii) the α of $\hat{k}_1$ and $\hat{k}_2$ are assigned the same value of $\alpha_{\hat{k}}$. A graphical summary of the splitting strategy is depicted on the right side of Fig. 3. Finally, note that – as for the pruning strategy – more than one topic can be splitted after a learning phase for speedup purposes.

4 Experimental Evaluation

In order to evaluate our robust initialization schemes, we performed several experiments on 5 different datasets. A summary of the employed datasets is

Table 2. Summary of the employed dataset. Columns W, D and Z correspond to the number of words, documents, and topics respectively.

Dataset name	References	Type of words	N	D	Z
1. HIV gag proteins	[24]	Protein sequence	1680	204	100
2. Lung genes	[2]	Genes	12600	203	100
3. FragBag	[29]	3D protein fragments	400	2928	100
4. Flickr images	[10]	Heterogeneous image features	82	60000	50
5. Science magazine	[23]	Textual words	24000	36000	100

reported on Table 2, where for each dataset we indicated its name, the number of words N (i.e. the dictionary size), the number of objects D, and the number of topics Z we employed for learning (when available, this number corresponds to the optimal choice found by the authors of the papers in the reported references).

We took these datasets from heterogeneous applicative contexts in the literature, which involve a wide variety of tasks, ranging from classification and clustering, to feature selection and visualization. Due to this heterogeneity, quantities such as the classification error can not be employed as a general measure of performance. Therefore, we resorted to two other validation indices: the first one is based on the log-likelihood of the learned model (on both the training set and an held out testing set), the second one takes into account the coherence of the learned topics. In both cases, we divided each dataset in a training and testing set using 10-fold crossvalidation, repeating the random subdivisions 3 times. For each fold and each repetition, we employed the proposed approaches to learn the LDA on the training set[2]. For the splitting approach, we set K_{small} to 2, and the Gaussian noise variance σ to 0.01. After a preliminary evaluation, we found that this noise parameter does not influence much results, provided that it is reasonably small (we found that performances deteriorate when $\sigma \geq 0.1$). For the pruning approach, we set K_{large} equal to the number of documents for the first three datasets, whereas we set it to 1000 for the Flickr images.

We compared our strategies with the random initialization (the currently most employed method), as well as with the technique proposed in [14], where the authors propose to initialize the β distribution by performing a Latent Semantic Analysis (LSA) on the training bag of words matrix: we will refer to this initialization technique as LSA. Please note that this method has been originally designed for initializing a slightly different topic model called PLSA. Its generalization to LDA is easy, because in PLSA the Dirichlet distribution is not employed, and θ is estimated point-wise (the equivalence between PLSA and LDA has been demonstrated in [18]): however, it is not clear how to initialize α. We decided to initialize $\alpha_k = 1\ \forall k$, this corresponding to a uniform prior over θ.

[2] We employed the public Matlab LDA implementation available at http://lear.inrialpes.fr/~verbeek/software.php.

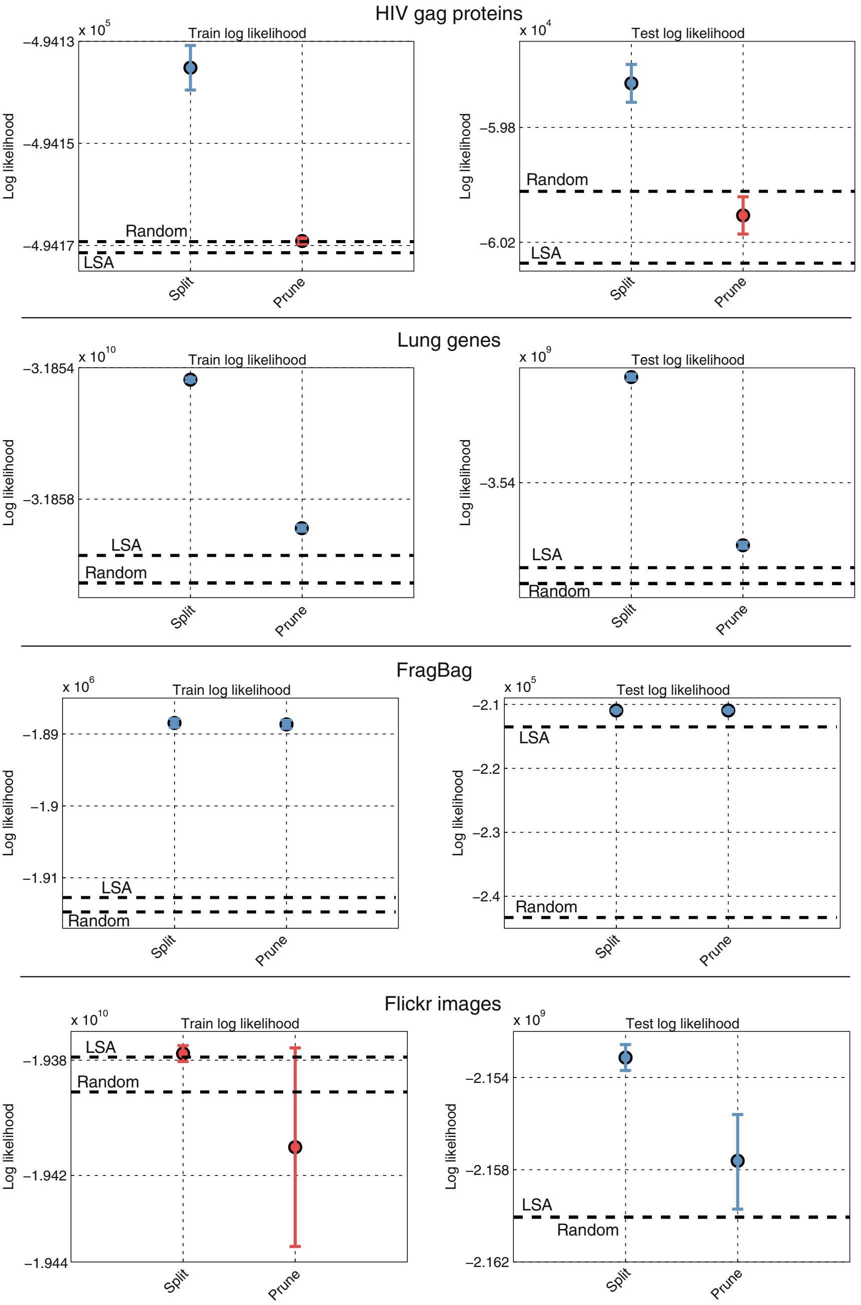

Fig. 4. Log-likelihood of the proposed methods for the different dataset. On the left, the log-likelihood of the training set. On the right, the log-likelihood evaluated on the held out testing set.

4.1 Log-Likelihood Evaluation

We firstly assessed the log-likelihood of the trained LDA models, on both a training and an held out testing set: while the log-likelihood of the training set indicates the quality of the learned model, the log-likelihood of the testing set gives insights into the generalization capability. Such log likelihoods, averaged over folds and repetitions, are shown in Fig. 4, for the first 4 datasets: the column on the left represents log-likelihoods obtained on the training set, whereas the column on the right depicts the ones obtained on the testing (held out) set. The dashed lines indicate the log-likelihood obtained with the Random and LSA methods we compared against, whereas the dots correspond to the log-likelihoods obtained with the proposed approaches. Finally, the bars correspond to the 95 % confidence intervals computed after a t-test, performed to assess if the results obtained with the proposed approaches led to a statistically significant improvement over the best-performing method (among the random and LSA initialization schemes – we highlighted statistically significant results in blue). From the figure it can be noted that the splitting scheme is on average the best one, being able to outperform other approaches in every case except one. The pruning scheme, even if reaching satisfactory results on 5 cases out of 8, seems to be slightly worse.

4.2 Coherence Evaluation

As a second measure of evaluation, we employed a measure of topic coherence to evaluate the proposed approaches. The coherence is essentially a score that is given to a single topic by measuring the degree of semantic similarity between highly probable words in the topic. Several coherence measures have been proposed in the past [22,32], and they are aimed at distinguish between topics that are semantically interpretable and topics that are artifacts of statistical inference. In this paper we adopted the internal criterion of Umass coherence [22]. We chose this in particular because it does not rely on an external corpus providing the ground-truth, which can be available in the text domain, but is absent in the other scenarios considered here. The Umass coherence defines a score based on a set of “topic words” V_k, which is created by retaining the top probable words in the topic (ranked by β probabilities). The Umass coherence of topic k is defined as

$$coherence(V_k) = \sum_{v_i, v_j \in V_k} score(v_i, v_j) \tag{5}$$

where

$$score(v_i, v_j) = \log \frac{p(v_i, v_j) + 1/D}{p(v_i)p(v_j)} \tag{6}$$

In the equation, $p(v_i, v_j)$ indicates the frequency of documents containing words v_i and v_j, and $p(v_i)$ measures the frequency of documents containing v_i. Note that the Umass computes these frequencies over the original corpus used to train the topic models: it attempts to confirm that highly probable words in the topic

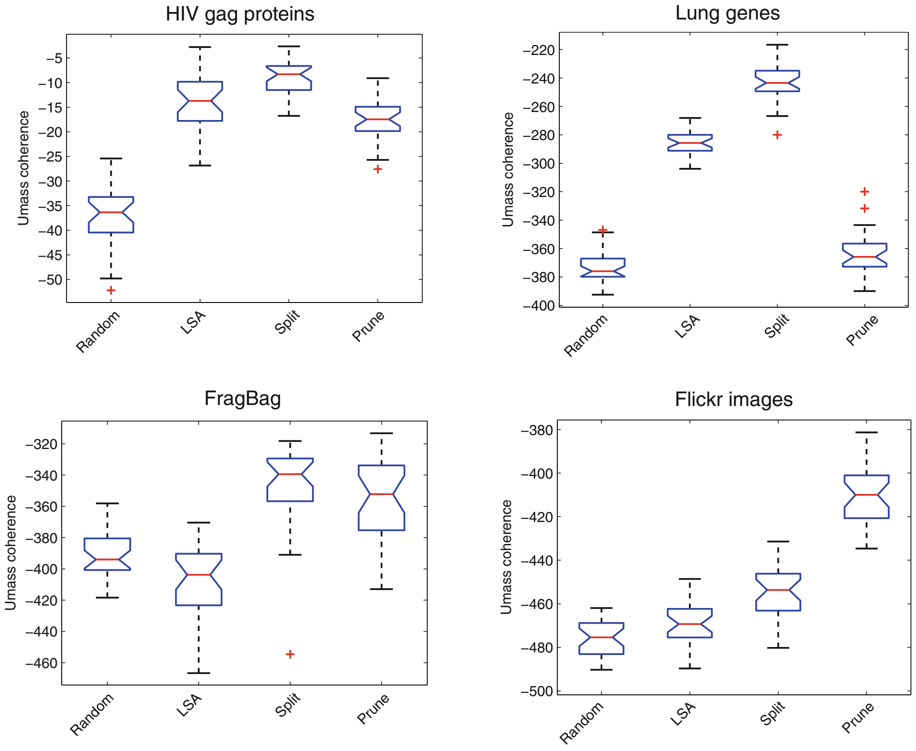

Fig. 5. Umass coherence for the different datasets and the different initialization schemes. The boxplot can be useful to assess statistical significance.

indeed co-occur in the corpus. In the end, the average coherence over the whole set of topics is reported as performance: a higher mean coherence indicates an overall better capability of the model to correctly cluster words in topics.

In our evaluation, for each fold and each repetition of each dataset, we applied the proposed approaches to learn the LDA on the training set. Then, as done before [32], we retained for each topic the top 10 words and computed the Umass coherence for all topics. Finally we averaged the coherences of all topics to get a final single score for the model. Coherence results, averaged over different folds and repetitions, are presented as boxplots in Fig. 5. Each box describes an evaluated method, and the red bar is the median coherence over the 30 repetitions (10 folds, randomly extracted 3 times). The edges of the blue box are the 25^{th} and 75^{th} percentiles, while the whiskers (black dashed lines) extend to the most extreme data points not considered outliers. Outliers are plotted individually with a red cross. Two medians are significantly different (at 95 % confidence) if their notches do not overlap. The splitting strategy always significantly outperforms the state of the art, thus confirming the suitability of this initialization strategy.

Concerning the pruning approach, we noticed that on the HIV and Lung datasets – while surpassing the random initialization – it is not competitive with respect to the other initialization techniques. On the contrary, on the last two datasets (FragBag and Flickr images), this strategy performs adequately well, achieving very high topic coherence on the Flickr images dataset in particular. Interestingly, we can observe that the HIV and lung datasets, due to the peculiar applicative scenario, present more words than objects, whereas the FragBag and Flickr images have a larger number of documents than words.

As a final consideration, we compared the computation times of the different initialization strategies. All the algorithms have been implemented in Matlab and run on a quad-core Intel Xeon E5440 @ 2.83GHz, with 4GB of RAM. The pruning strategy requires the largest running time, several order of magnitude greater than the other strategies. For what concerns the other strategies, it should be observed that in general results depend on the characteristics of the dataset (number of documents and number of words). In fact, when the number of documents D is fairly small (as for the HIV gag dataset), the running times of the LSA and splitting strategies are comparable with the random one: even if initializing the parameters α and β at random is almost istantaneous, more iterations are required to achieve convergence in the learning phase. For example, learning the LDA model starting from one random initialization required 158 iterations, starting from the LSA initialization required 140 iterations and starting from the splitting initialization required 134. On the contrary, when the number of documents is really high (as for the Psychoflickr dataset), then the random initialization is approximately 5 times faster. However, it may still be reasonable to raise the computational burden and adopt the splitting strategy, motivated by the quality of the solution that can be achieved (in many cases, the learning is done only once, off-line).

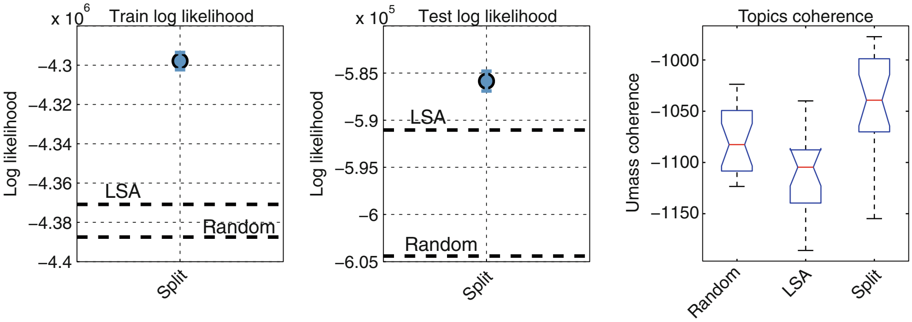

Fig. 6. Science magazine results. The first two panels depict as a dot the train and test log likelihood of the splitting strategy, which is always significantly higher than the dashed lines, corresponding to the random and LSA initialization techniques. On the right, comparison between the Umass coherence of the different approaches.

4.3 Science Magazine Dataset

An important consideration that has to be made for the pruning strategy is that, although it seems suited in several situations, it is not applicable when the number of documents is very high. This is the case of the Science magazine dataset, which we discuss separately because we evaluated only the splitting strategy. Results on this dataset are reported on Fig. 6. Also in this case, it can be noted that the splitting strategy reaches satisfactory log-likelihood values, as well as coherence scores, when compared with the other alternatives.

5 Conclusions

In this paper we proposed two novel strategies to initialize the Latent Dirichlet Allocation (LDA) topic model, that aim at fully exploiting the characteristics of the model itself. The key idea is to employ a splitting or a pruning approach, where each training session is initialized from an informative situation derived from the previous training phase. Then, in order to choose the best topic to split/prune, we leveraged the intrinsic information derived from the model: in particular, we exploit the parameter α of the Dirichlet distribution, that can be seen as a measure of the prior "importance" of each topic. The quality of the LDA model learned using our approaches has been experimentally evaluated on 5 different datasets, taken from heterogeneous contexts in the literature. Results suggested that the splitting and pruning strategies are well suited, and can boost the model in terms of its train and test log likelihood, as well as in terms of the coherence of the discovered topics.

References

1. Asuncion, H., Asuncion, A., Taylor, R.: Software traceability with topic modeling. In: Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering, ICSE 2010, vol. 1, pp. 95–104 (2010)
2. Bhattacharjee, A., Richards, W., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E., Lander, E., Wong, W., Johnson, B., Golub, T., Sugarbaker, D., Meyerson, M.: Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. **98**(24), 13790–13795 (2001)
3. Bicego, M., Lovato, P., Perina, A., Fasoli, M., Delledonne, M., Pezzotti, M., Polverari, A., Murino, V.: Investigating topic models' capabilities in expression microarray data classification. IEEE/ACM Trans. Comput. Biol. Bioinform. **9**(6), 1831–1836 (2012)
4. Bicego, M., Murino, V., Figueiredo, M.: A sequential pruning strategy for the selection of the number of states in hidden Markov models. Pattern Recogn. Lett. **24**(9), 1395–1407 (2003)
5. Blei, D.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
6. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

7. Boley, D.: Principal direction divisive partitioning. Data Mining Knowl. Disc. **2**(4), 325–344 (1998)
8. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)
9. Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., Blei, D.: Reading tea leaves: how humans interpret topic models. Adv. Neural Inf. Process. Syst. **22**, 288–296 (2009)
10. Cristani, M., Vinciarelli, A., Segalin, C., Perina, A.: Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 213–222 (2013)
11. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Ser. B (Methodol.) **39**, 1–38 (1977)
12. Dhillon, I., Mallela, S., Kumar, R.: A divisive information theoretic feature clustering algorithm for text classification. J. Mach. Learn. Res. **3**, 1265–1287 (2003)
13. Elidan, G., Friedman, N.: The information bottleneck EM algorithm. In: Proceedings of the Uncertainty in Artificial Intelligence, pp. 200–208 (2002)
14. Farahat, A., Chen, F.: Improving probabilistic latent semantic analysis with principal component analysis. In: EACL (2006)
15. Fayyad, U., Reina, C., Bradley, P.: Initialization of iterative refinement clustering algorithms. In: Knowledge Discovery and Data Mining, pp. 194–198 (1998)
16. Figueiredo, M.A.T., Leitão, J.M.N., Jain, A.K.: On fitting mixture models. In: Hancock, E.R., Pelillo, M. (eds.) EMMCVPR 1999. LNCS, vol. 1654, pp. 54–69. Springer, Heidelberg (1999)
17. Frey, B., Jojic, N.: A comparison of algorithms for inference and learning in probabilistic graphical models. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1–25 (2005)
18. Girolami, M., Kabán, A.: On an equivalence between plsi and lda. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Formaion Retrieval, pp. 433–434 (2003)
19. Hazen, T.: Direct and latent modeling techniques for computing spoken document similarity. In: 2010 IEEE Spoken Language Technology Workshop (SLT), pp. 366–371 (2010)
20. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. **42**(1–2), 177–196 (2001)
21. Lienou, M., Maitre, H., Datcu, M.: Semantic annotation of satellite images using Latent Dirichlet Allocation. IEEE Geosci. Remote Sens. Lett. **7**(1), 28–32 (2010)
22. Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 262–272 (2011)
23. Perina, A., Kim, D., Turski, A., Jojic, N.: Skim-reading thousands of documents in one minute: data indexing and visualization for multifarious search. In: Workshop on Interactive Data Exploration and Analytics, IDEA 2014 at KDD (2014)
24. Perina, A., Lovato, P., Jojic, N.: Bags of words models of epitope sets: HIV viral load regression with counting grids. In: Proceedings of International Pacific Symposium on Biocomputing (PSB), pp. 288–299 (2014)
25. Quinn, K., Monroe, B., Colaresi, M., Crespin, M., Radev, D.: How to analyze political attention with minimal assumptions and costs. Am. J. Polit. Sci. **54**(1), 209–228 (2010)

26. Roberts, S., Husmeier, D., Rezek, I., Penny, W.: Bayesian approaches to gaussian mixture modeling. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1133–1142 (1998)
27. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York (1986)
28. Segalin, C., Perina, A., Cristani, M.: Personal aesthetics for soft biometrics: a generative multi-resolution approach. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 180–187 (2014)
29. Shivashankar, S., Srivathsan, S., Ravindran, B., Tendulkar, A.: Multi-view methods for protein structure comparison using Latent Dirichlet Allocation. Bioinformatics **27**(13), i61–i68 (2011)
30. Smaragdis, P., Shashanka, M., Raj, B.: Topic models for audio mixture analysis. In: NIPS Workshop on Applications for Topic Models: Text and Beyond, pp. 1–4 (2009)
31. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD workshop on text mining. vol. 400, pp. 525–526 (2000)
32. Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D.: Exploring topic coherence over many models and many topics. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 952–961 (2012)
33. Ueda, N., Nakano, R.: Deterministic annealing EM algorithm. Neural Netw. **11**(2), 271–282 (1998)
34. Wang, C., Blei, D., Li, F.F.: Simultaneous image classification and annotation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1903–1910 (2009)
35. Wu, C.: On the convergence properties of the EM algorithm. Ann. Stat. **1**(1), 95–103 (1983)
36. Yang, S., Long, B., Smola, A., Sadagopan, N., Zheng, Z., Zha, H.: Like like alike: joint friendship and interest propagation in social networks. In: Proceedings of the 20th International Conference on World Wide Web (WWW), WWW 2011, pp. 537–546 (2011)