Boosting SpLSA for Text Classification

Julio Hurtado[1], Marcelo Mendoza[2(✉)], and Ricardo Ñanculef[2]

[1] Pontificia Universidad Católica de Chile, Santiago, Chile
julio.hurtado@puc.cl
[2] Universidad Técnica Federico Santa María, Valparaíso, Chile
{marcelo.mendoza,ricardo.nanculef}@usm.cl

Abstract. Text classification is a challenge in document labeling tasks such as spam filtering and sentiment analysis. Due to the descriptive richness of generative approaches such as probabilistic Latent Semantic Analysis (pLSA), documents are often modeled using these kind of strategies. Recently, a supervised extension of pLSA (spLSA [10]) has been proposed for human action recognition in the context of computer vision. In this paper we propose to extend spLSA to be used in text classification. We do this by introducing two extensions in spLSA: (a) Regularized spLSA, and (b) Label uncertainty in spLSA. We evaluate the proposal in spam filtering and sentiment analysis classification tasks. Experimental results show that spLSA outperforms pLSA in both tasks. In addition, our extensions favor fast convergence suggesting that the use of spLSA may reduce training time while achieving the same accuracy as more expensive methods such as sLDA or SVM.

Keywords: Probabilistic Latent Semantic Analysis · Model Regularization · Label uncertainty · Text classification

1 Introduction

Given the large amount of data held in text format, modeling topics in text has been of growing interest in the last decade. Generative strategies based on probabilistic Latent Semantic Analysis (pLSA [6]) provide a solid theoretical base and a flexible framework for this task that allow for the modeling of various kind of corpora. The idea driving these models consists in introducing a set of latent variables that allow one to discover relationships between terms. These kind of strategies are known as topic models.

Topic models are fundamentally divided into two broad approaches: techniques derived from pLSA, which introduce latent variables without assuming distributional priors, and techniques based on Latent Dirichlet Allocation (LDA [2]) which assumes Dirichlet priors on topics and vocabularies.

From a machine learning perspective, topic models such as pLSA, correspond to unsupervised learning systems capable to discover structures in data without supervision. At the other end we have supervised learning systems, which are able to exploit annotated examples to predict future annotations. Classification techniques like the Multinomial Naive Bayes model [7] and the Support Vector

C. Beltrán-Castañón et al. (Eds.): CIARP 2016, LNCS 10125, pp. 142–149, 2017.
DOI: 10.1007/978-3-319-52277-7_18

Machine (SVM [4]) are well-known examples of supervised systems with many applications in text mining, including e.g. spam filtering [1] and sentiment detection [8]. Unfortunately, though often very accurate, these models either lack the descriptive richness or the efficiency of text-oriented methods such as pLSA.

Recently, a supervised version of pLSA (spLSA [10]) has been proposed to tackle the problem of human action recognition in computer vision. By adding labels into the generative process, spLSA was endowed with more discriminative power for classification tasks when annotations are available. Surprisingly, spLSA has not yet been evaluated in text classification despite the fact that pLSA was originally designed for text modeling. To the best of our knowledge, in this paper we present the first evaluation of spLSA in text classification tasks. In addition, we introduce two extensions to spLSA: (a) Regularized spLSA, a variant that introduces label co-variance minimization into the Expectation-Maximization (EM) model fitting algorithm, and (b) Label uncertainty, an extension that allow us to deal with noisy labels, a common problem in human and machine annotated corpora [11].

2 Related Work

Due to its simplicity, the Multinomial Naive Bayes classifier (MNB [7]) is one of the most used methods for text classification. MNB assumes (class) conditional independence among terms, which reduces the complexity of the model in terms of the number of parameters and makes its estimation from data more efficient and reliable. However, in sparse data sets its performance tends to decrease. LDA-based methods introduce smoothing over the data set using Dirichlet priors [2], favoring classification tasks over sparse data sets. The supervised extension of LDA (sLDA [3]) shows good results in classification tasks but introduces difficulties in parameter tuning. In fact, due to the use of Dirichlet priors, the algorithm needs to tune distributional hyper-parameters. The lack of clear procedures for tuning is a drawback of this approach.

Discriminative approaches such as the Support Vector Machine and Logistic Regression [4] are also used for these tasks. In general, these approaches outperform generative approaches in terms of classification accuracy. However, they are less used due to difficulties associated with vocabulary characterization. In practice, generative approaches offer advantages regarding corpus descriptiveness, favoring tasks such as content analysis and term indexing, which are key tasks in information extraction and document processing.

3 Supervised pLSA and Our Extensions

3.1 Variables and Assumptions

An observation in a labeled corpus $\mathcal{D}$ is the realization of three observed random variables: A document ($\mathbf{d}$), a bag-of-words ($\mathbf{w}$) which describes the content of $\mathbf{d}$, and a label ($\mathbf{y}$) which indicates the membership of $\mathbf{d}$ to a given set of categories

Λ. The probability of observing a given realization of these variables $\langle \mathbf{d} = d, \mathbf{w} = w, \mathbf{y} = y \rangle$ is denoted by $P(d, w, y)$.

A document d is defined as a sequence of words selected from a vocabulary V. As documents share words, several semantic relationships among documents arise. By modeling hidden relationships between document and terms through latent variables, a latent semantic representation of the corpus is built.

While labels are typically noisy, words and documents convey information. By introducing latent factors, labels can be stressed by words and documents, allowing to model label uncertainty, discarding rare, unexpected labels and bearing out only data supported labels.

3.2 Supervised pLSA (spLSA)

The idea of a topic model is that documents are represented by mixtures of memberships over a latent space. Basically, spLSA considers a label as another random variable. Then, a generative approach where every observed variable depends on latent factors is used:

$$P_z(d, w, y) = \sum_z P(w|z, y, d) \cdot P(z, y, d) = \sum_z P(w|z) \cdot P(y, d|z) \cdot P(z)$$
$$= \sum_z P(w|z) \cdot P(y|z) \cdot P(d|z) \cdot P(z).$$

The last expression corresponds to a generative process of $\mathcal{D}$ from the model.

3.3 Model Fitting

Model parameters (θ) can be determined by maximizing the log-likelihood function of the generative process:

$$\mathcal{L}_\theta = \sum_{y \in \Lambda} \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{V}} n(d, w, y) \log P(d, w, y),$$

where $n(d, w, y)$ denotes the number of occurrences of (d, w, y) in $\mathcal{D}$. The Expectation Maximization algorithm (EM) is the standard procedure for parameter estimation in latent variable models. The E-step of the algorithm estimates $\hat{P}(z|d, w, y)$ from the generative model. Then, the M-step estimates the generative model components ($\hat{P}(w|z)$, $\hat{P}(y|z)$, $\hat{P}(d|z)$, and $\hat{P}(z)$) by likelihood marginalization.

3.4 Label Inference on New Documents

Document labeling using spLSA is straightforward. Starting from a uni-gram language model for a new document d with n_d terms, we have

$$P(d|z) \propto \prod_{i=1}^{n_d} P(w_i|z),$$

and then, by applying $P(y, d) = \sum_z P(y, d, z) = \sum_z P(y|z) \cdot P(d|z) \cdot P(z)$,

$$P_z(y = l, d) \propto \sum_{z \in \mathcal{Z}} \left[\prod_{i=1}^{n_d} P(w_i|z) \cdot P(y = l|z) \cdot P(z) \right].$$

Finally, d is labeled with $y = l$ if $P(y = l|d) \propto P(y = l, d)$ is greater than any other $P(y = \Lambda \backslash l|d)$.

3.5 A First Extension: SpLSA Regularization

Since the objective function in spLSA is not convex, the EM-based iterative method will tend to converge to local optima. As observed from the introduction of pLSA [6], many of these local optima are plagued by over-fitting. Therefore, it is advisable to endow the algorithm with a regularization technique. In this paper, we propose the incorporate additive Tikhonov-based regularizers [9] in the M-step of the model fitting phase. Concretely, we introduce regularization for latent factor co-variance minimization.

Our idea is to penalize solutions where the confusion matrix ($\hat{P}(y \mid z)$) is far from the identity $\mathcal{I}$. We do this by minimizing the co-variance between $P(y = l|z)$ and $P(y = l|Z \backslash z)$, forcing a match between each latent factor with only one label. Note that this procedure is equivalent to the diagonalization of the confusion matrix. The latent factor co-variance minimization modifies the likelihood function as follows:

$$\mathcal{L} = \mathcal{L}_\theta - \sum_{l \in \Lambda} \text{Cov}\left(\hat{P}(y = l|z), \hat{P}(y = l|z^{'} \in Z \backslash z) \right).$$

The regularized maximum likelihood function can be maximized by modifying the M-step, where label probabilities conditioned on latent factors are estimated as in Sect. 3.3 with an additional step:

$$\hat{P}(y = l|z) = \hat{P}(y = l|z) \cdot \left[1 - \sum_{z^{'} \in Z \backslash z} \hat{P}(y = l|z^{'}) \right].$$

As a consequence, the EM-algorithm will penalize $\hat{P}(y = l|z)$ estimates for labels highly correlated to other latent variables. Note that $\hat{P}(y = l|z)$ is maximized when $(1 - \sum_{z^{'} \in Z \backslash z} \hat{P}(y = l|z^{'})) = 1$, which it is precisely what we are looking for.

Finally, we note that the co-variance regularizer may be applied to term probabilities or document probabilities, in a similar fashion. All these variants will be evaluated in our experiments.

3.6 A Second Extension: Modeling Label Uncertainty

Label uncertainty is common in text mining applications where annotations are obtained via crowd-sourcing or distant supervision [11]. To handle label uncertainty in spLSA, we modify the label estimation procedure by introducing

the possibility to flip the label available in the data set. Formally, let $b \in [0, 1]$ the flipping probability parameter. To consider both possible labels (binary case), we introduce a convex combination conditioned on b in the M-step for $\hat{P}_b(y|z)$ as follows:

$$\hat{P}_b(y|z) = \frac{1}{Q_z} \cdot [(1-b) \cdot \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{V}} n(d, w, y) \cdot \hat{P}(z|d, w, y = l) + b \cdot \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{V}} n(d, w, y) \cdot \hat{P}(z|d, w, y = \Lambda \backslash l)],$$

where l is the label provided by the data set. Note that b controls the level of uncertainty by blending $\hat{P}(y = l|z)$ and $\hat{P}(y = \Lambda \backslash l|z)$ into a unified model. When $b = 0$ the equation is equivalent to $\hat{P}(y = l|z)$ and when $b = 1$ is equivalent to $\hat{P}(y = \Lambda \backslash l|z)$. Thus, values between 0 and 1 will increase/decrease the confidence on the label provided by the data set. When $b = 0.5$ we give to the label a maximum degree of uncertainty, being both options (l or $\Lambda \backslash l$) equally probable.

Our method can be extended to the multi-class scenario, by applying a one-versus-the-rest strategy for $\hat{P}_b(y|z)$ estimation. In this case, it is enough to assume that all the labels that belong to $\Lambda \backslash l$ are equally probable.

4 Experiments

To evaluate our proposal, we performed experiments on two text labeled data sets. The first was designed for email spam filtering tasks [1]. We use this data set to illustrate the impact of regularization on text classification. The second data set contains tweets labeled as positive or negative regarding emotion polarity. It comprises a number of tweets labeled using distant supervision (it uses emoticons to automatically label each tweet) and thus it can be regarded as a data set with noisy labels. We use this data set to assess our strategy for handling label uncertainty.

4.1 Data Sets

Below, we provide a brief description of the data sets used in our experiments.

- **Email spam data set** [1]: This data set comprises a 700-email subset for training and a 260-email subset for testing. Both the training and testing subsets contain 50% spam messages and 50% non-spam messages. The data set comprises a vocabulary of 2.500 words. The data set is available in: http://csmining.org/index.php/ling-spam-datasets.html
- **Twitter distant supervision data set** [5]: It includes 1.600.000 tweets written in English, labeled as positive or negative regarding emotion polarity, inferring labels from emoticons. A set of 359 manually labeled tweets is provided for validation purposes (177 as negative and 182 as positive). The vocabulary is compounded by 267.013 terms. The data set is available in: http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip

4.2 Results

We start by exploring the performance of spLSA and co-variance regularization in text classification. We compare our proposals with pLSA [6], sLDA [3], Multinomial Naive Bayes (MNB [7]) and Support Vector Machines (SVMs [4]). For SVM, we used a C-SVM formulation provided by Liblinear [4], with a tuning over C based on five-fold cross validation ($C = 0.1$ in both data sets). For sLDA, we used default values for hyper-parameters[1]. We tested three variants of regularized spLSA, obtained by applying co-variance regularization on terms ($\hat{P}(w|z)$), documents ($\hat{P}(d|z)$), and labels ($\hat{P}(y|z)$). As both data sets rely on binary classes, we used two latent factors for the experiments (i.e. $T = 2$ for sLDA and pLSA).

We assess all the methods in terms of test accuracy and running time. In the first data set, experiments were conducted using a tolerance value of $1E{-}08$ for early stopping to guarantee local convergence (likelihood at single precision is enough for this experiment due to the small size of the data set). Training and testing performance are summarized in Table 1.

As Table 1 shows, the four variants of spLSA get good results in text classification. A slight difference in favor of SVM is observed regarding testing accuracy. As expected, training accuracies are better than testing accuracies. Note that the training times of spLSA are less than those incurred by pLSA, suggesting that spLSA helps to obtain fast convergence. In fact, the fastest solution is achieved using label regularization. Note also that spLSA and sLDA find the same solution, but spLSA is faster by one order of magnitude. In addition, this solution is better in terms of accuracy than the one achieved using MNB. In summary, results on this data set indicate that the use of regularization on spLSA reduces the number of iterations required for convergence without compromising classification accuracy.

Table 1. Classification accuracy performance on the email spam data set.

Method	Variant	Training acc.	Testing acc.	Time[s]
MNB	-	0.975	0.946	0.24
SVM	-	0.975	0.973	0.47
sLDA	-	0.975	0.965	3.12
pLSA	-	0.958	0.946	0.97
spLSA	-	0.975	0.965	0.24
spLSA	Reg. $\hat{P}(w\|z)$	0.975	0.965	0.24
spLSA	Reg. $\hat{P}(d\|z)$	0.975	0.965	0.24
spLSA	Reg. $\hat{P}(y\|z)$	0.975	0.965	0.19

[1] $\alpha = \frac{50}{T}$, and $\beta = \frac{200}{W}$, T is the number of topics and W the vocabulary size.

Table 2. Classification accuracy performance on the Twitter distant supervision data set. Bold fonts indicate the best performance result.

Method	Variant	Training acc.	Testing acc.	Time[s]
MNB	-	0.797	0.777	312
SVM	-	0.817	0.816	624
sLDA	-	0.797	0.739	1245
pLSA	-	0.561	0.565	799
spLSA	-	0.817	0.777	326
spLSA	Reg. $\hat{P}(w\|z)$	0.817	0.777	323
spLSA	Reg. $\hat{P}(d\|z)$	0.817	0.777	331
spLSA	Reg. $\hat{P}(y\|z)$	0.817	0.793	81

In the second data set, we used a tolerance value of $1E{-}16$ for early stopping (likelihood at double precision due to the size of the data set). Training and testing performance results are summarized in Table 2. These results show that spLSA outperforms pLSA by more than 20% accuracy points in its four variants, which, in turn, achieve a similar level of accuracy. As expected, the SVM is slightly more accurate at the expense of a greater computational cost ($\sim 2\times$ or more). Label regularization favors a significantly faster convergence, reducing training time to 81 secs with a testing accuracy better than the ones achieved by sLDA and MNB.

Finally, we analyze the impact of label uncertainty in the Twitter distant supervision data set, varying the value of the flipping probability parameter b in the range 0 to 0.6, with increments of 0.01. We use the label regularization variant for fast convergence. These results are shown in Fig. 1. The figure shows,

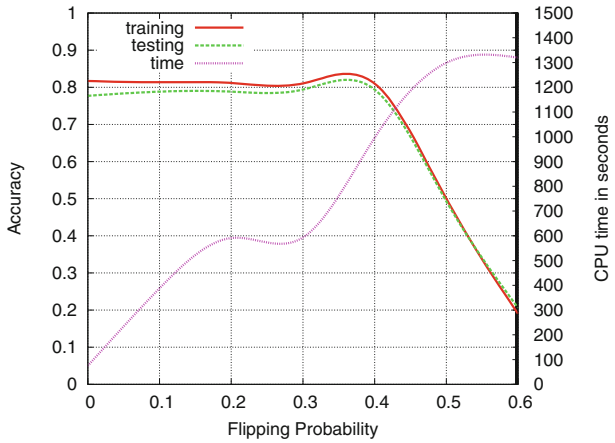

Fig. 1. Training and testing accuracy for different levels of label uncertainty.

the accuracy achieves values around 80% in both data partitions, with a small difference in favor of the training part of the data. This difference is reduced when b increases, until accuracy achieves its maximum for b in the range 0.3–0.4. As expected, the performance dramatically decreases when b tends to 0.5. In addition, we can observe high computational costs for high values of b. The best performance is achieved around $b = 0.35$, where our proposal reaches the same performance achieved by SVM.

5 Conclusions

Two extensions of spLSA for text classification has been proposed. Experimental results show that the both methods are feasible, achieving very competitive results in terms of accuracy. Two main findings arise. First, label regularization in spLSA allows to obtain faster convergence and thus lower training times. Second, handling label uncertainty in spLSA allows improvements in terms of test accuracy but increases computational costs.

References

1. Androutsopoulos, I., Koutsias, J., Chandrinos, K., Paliouras, G., Spyropoulos, C.: An evaluation of naive bayesian anti-spam filtering. In: 11th ECML, Workshop on Machine Learning in the New Information Age, pp. 9–17 (2000)
2. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Mach. Learn. Res. (JMLR) **3**(4–5), 993–1022 (2003)
3. Blei, D., McAuliffe, J.: Supervised topic models. In: Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems (NIPS), Vancouver, British Columbia, Canada, pp. 121-128 (2007)
4. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: A library for large linear classification. J. Mach. Learn. Res. (JMLR) **9**, 1871–1874 (2008)
5. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project report, Stanford University (2009)
6. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. **42**(2), 177–196 (2001)
7. McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. In: Proceedings of the International Conference on Machine Learning (ICML), Madison, WI, USA, pp. 41–48 (1998)
8. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends in Inf. Retrieval **2**(1–2), 1–135 (2008)
9. Tikhonov, A., Arsenin, V.: Solutions of Ill-posed Problems. Winston & Sons, Great Falls (1977)
10. Wang, T., Liu, C.: Human action recognition using supervised pLSA. Int. J. Signal Process. Image Process. Pattern Recognit. **6**(4), 403–414 (2013)
11. Frenay, B., Verleysen, M.: Classification in the presence of label noise: a survey. IEEE Trans. Neural Netw. Learn. Syst. **25**(5), 845–869 (2014)