Semi-supervised Online Kernel Semantic Embedding for Multi-label Annotation

Jorge A. Vanegas[1(✉)], Hugo Jair Escalante[2], and Fabio A. González[1]

[1] MindLab Research Group, Universidad Nacional de Colombia, Bogotá, Colombia
{javanegasr,fagonzalezo}@unal.edu.co
[2] Instituto Nacional de Astrofísica, Optica y Electrónica (INAOE), Puebla, Mexico
hugojair@ccc.inaoep.mx

Abstract. This paper presents a multi-label annotation method that uses a semantic embedding strategy based on kernel matrix factorization. The proposed method called Semi-supervised Online Kernel Semantic Embedding (SS-OKSE) learns to predict the labels of a document by building a semantic representation of the document features that takes into account the labels, when available. A remarkable characteristic of the algorithm is that it is based on a kernel formulation that allows to model non-linear relationships. The SS-OKSE method was evaluated under a semi-supervised learning setup for a multi-label annotation task, over two text document datasets and was compared against several supervised and semi-supervised methods. Experimental results show that SS-OKSE exhibits a significant improvement, showing that a better modeling can be achieved with an adequate selection/construction of a kernel input representation.

Keywords: Semantic representation · Semi-supervised learning
Learning on a budget · Multi-label annotation

1 Introduction

The automatic multi-label annotation problem presents several applications in areas as diverse as text and music categorization and classification, semantic labeling of images and videos, medical diagnosis, and functional genomics, among others [11]. Several methods transform the problem of multi-label learning to a conventional classification problem (i.e., a set of binary classification problems solved independently). Unfortunately this kind of approaches present two principal drawbacks, first, usually they do not scale well when the number of labels and/or instances increase, and second, these approaches do not have into account the possible strong correlations between the labels. Another important issue is that these approaches require a significant amount of labeled data to achieve a reasonable generalization performance. In multi-label learning, this issue is more evident than in single-class classification since manually assigning multiple labels

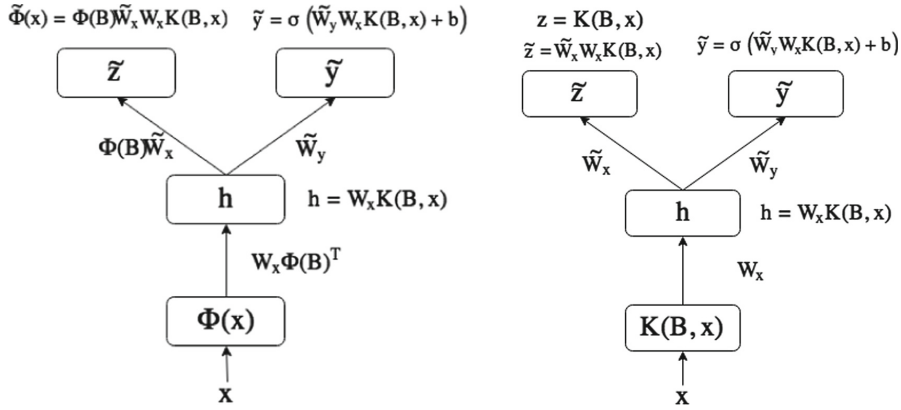
M. Mendoza and S. Velastín (Eds.): CIARP 2017, LNCS 10657, pp. 693–701, 2018.
https://doi.org/10.1007/978-3-319-75193-1_83

is more time-demanding than assigning unique global labels. A possible strategy to deal with a large number of labels and the lack of annotated instances is to find a compact representation of them by using, for instance, a dimensionality reduction method. This approach is followed by multi-label latent space embedding methods, which have recently shown competitive results. There are several strategies to construct the latent semantic space, and most of them proposed supervised and semi-supervised extensions. The important thing about this kind of methods, is that this compact semantic representation also can be modeled by the discriminative structure of not only labeled but also unlabeled data. In this paper, we present a method for multi-label annotation based on semantic embedding that finds a common semantic space based on the kernel feature representation of an instance and its corresponding labels that model a mapping between the feature representation and the annotation labels. The proposed method has three important characteristics: (1) the method is formulated as a semi-supervised learning algorithm that learns to construct a common semantic representation not only from labeled instances but also from unlabeled ones, (2) despite being based on kernels, the method scales well to deal with large datasets thanks to a budget restriction which allows tackling one of the main problems of kernel-based methods, that is the scalability in terms of number of training instances, and (3) the method is formulated as an on-line learning algorithm, based on stochastic gradient descent, which allows it to deal with large collections of data, achieving a significant reduction in memory requirements and computational load. Additionally, the latter characteristic allows the efficient implementation of the method in dataflow GPU frameworks such as Theano and TensorFlow, which are used for efficient training and simulation of deep neural networks.

The rest of this paper is organized as follows: Sect. 2 discusses the related work; Sect. 3 formally introduce the details of the proposed multi-label annotation method; Sect. 4 presents the experimental evaluation; and, finally, Sect. 5 presents some concluding remarks.

2 Multi-label Annotation Based on Semantic Embedding Methods

The existing methods for multi-label classification problems can be grouped into two main categories [11]: (1) problem transformation methods, which transform the multi-label classification problem into several single-label classification or regression problems and (2) algorithm adaptation methods, which extend specific learning algorithm in order to handle multi-label data directly. In the group of algorithm adaptation methods, we can find several adaptations to classical discriminative methods. For instance, Andrews et al. [1] proposed two extensions for Support Vector Machine (SVM) for multi-instance learning methods the miSVM and MISVM. miSVM treats instance labels as unobserved variables and maximizes the margin on instances. MISVM, in contrast, define a new concept of bag margin maximization that maximizes the difference between individual patterns. Unfortunately, these methods present two main problems: first,



(a) Conceptual model (b) Implemented model

Fig. 1. Conceptual model and the actual implementation of SS-OKSE.

Kernel-based methods usually do not scale well due to the high computational complexity caused by the kernel matrix that grows quadratically with the number of training instances, and second, these methods cannot be extended for a semi-supervised learning and require ground-truth labels for all training documents. Topic models can overcome the second problem by modeling a compact semantic representation that can be modeled only by the discriminative structure of the data. Classical semantic embedding models are usually unsupervised, but several extensions to add supervision have been proposed, for instance, several matrix factorization based methods have been extended to improve the semantic representation by taking advantage of label information [2,7]. Under the same approach, several probabilistic topic models have been extended. For instance, many works have extended the classical Latent Dirichlet Allocation (LDA) [4] to add supervised and semi-supervised information, such as Semi-supervised LDA [8], Maximum Entropy Discrimination Latent Dirichlet Allocation (MedLDA) [12], Partially Labeled LDA (PLLDA) and more recently the Semi-supervised Multi-label Topic Model (MLTM) [9]. Probabilistic topic models like LDA have the advantage of incorporating prior knowledge to guide the topic modeling process to improve both the quality of the resulting topics and of the topic labeling, but unfortunately are very computational demanding, making them prohibited for large scale problems.

In this paper we propose the Semi-supervised Online Kernel Semantic Embedding (SS-OKSE) method that is formulated for large scale problems by tackling two main issues: first, the method presents a budget restriction that reduces the computational complexity caused by the kernel matrix, and second, the proposed method is formulated as an on-line learning algorithm which allows it to deal with large collections of data.

3 Semi-supervised Online Kernel Matrix Factorization for Multi-label Annotation

In this paper we propose a multi-label latent space embedding method that constructs an intermediate semantic space modeled by the input representation projected in a feature space generated by a kernel function, this strategy can be seen as a matrix factorization problem in the kernel feature space ($\Phi(X) \simeq F_\Phi H$), where $X \in \mathbb{R}^{n \times l}$ describes the entire collection composed by l elements represented by $n-$dimensional vectors, $F_\Phi \in \mathbb{R}^{n \times r}$ is the basis matrix and $H \in \mathbb{R}^{r \times l}$ is the encoding matrix that represents all the input instances in a low $r-$dimensional space. Unfortunately, the calculation of this factorization is infeasible due to F_Φ depends explicitly on the mapping function $\phi(\cdot)$ (a mapping to a very highly dimensional space or even to an infinite-dimensional space). Therefore, instead of calculating directly F_Φ, we impose the restriction that the column vectors of F_Φ lie within the space of $\Phi(X)$, this is, F_Φ is composed of linear combinations of the X points in the feature space ($F_\Phi = \Phi(X)\tilde{W}_x$).

$$\Phi(X) \simeq \Phi(X)\tilde{W}_x H, \quad \Phi(X) \approx \Phi(B)\tilde{W}_x H \tag{1}$$

This restriction avoid the necessity of evaluating the data in the feature space, additionally, only a reduced number $b \ll l$ of representative instances are used to construct the basis matrix (we construct a budget kernel matrix $B \in \mathbb{R}^{b \times l}$ instead of the full kernel matrix $X \in \mathbb{R}^{l \times l}$) This mitigates the high computational cost of constructing the huge kernel matrix. In this paper, we propose not to construct an explicit representation in the semantic space but learn a mapping from the feature representation to this semantic space, and again, using only the b representative points to model the restricted kernel feature space ($H = W_x \phi(B)^T \phi(X) = W_x K(B, X)$). In this manner, the model learns two transformations what allows to project the original data representation to the lower-dimensional semantic space and at the same time to reconstruct from this semantic representation the original data in the feature space.

$$\Phi(X) \approx \Phi(B)\tilde{W}_x W_x K(B, X) \tag{2}$$

Additionally to the original feature representation, we want the semantic representation to also lie the label representation $Y \in \mathbb{R}^{m \times k}$, where m is the total number of possible labels and k a reduced number of annotated instances (i.e., $k \ll l$), as follows:

$$Y \approx \sigma\left(\tilde{W}_y H\right) = \sigma\left(\tilde{W}_y W_x K(B, X)\right) \tag{3}$$

where $W_y \in \mathbb{R}^{m \times r}$ is another transformation matrix that projects from the semantic representation to the label space, and finally, an additional non-linear function σ is used to add more flexibility to the model. Putting all these restriction together, the final model can be represented as is shown in Fig. 1a.

Loss function. The final loss function to be minimized forces the feature reconstruction by defining a squared minimum error and learns the binary label reconstruction by imposing a binary cross entropy function:

$$\begin{aligned}\min_{W_x,\tilde{W}_x,W_y} J_i\left(W_x,\tilde{W}_x,W_y\right) = \frac{\alpha}{2}\left\|\Phi\left(x_i\right)-\Phi\left(B\right)\tilde{W}_x W_x K\left(B,x_i\right)\right\|_F^2 \\ +\beta\sum_{i=0}^{k}\log\left(1+\exp\left(-y\cdot\tilde{W}_y W_x K\left(B,x_i\right)\right)\right) \\ +\frac{\lambda_1}{2}\left\|W_x\right\|_F^2+\frac{\lambda_2}{2}\left\|\tilde{W}_x\right\|_F^2+\frac{\lambda_3}{2}\left\|\tilde{W}_y\right\|_F^2 \qquad (4)\end{aligned}$$

where α and β control the relative importance of reconstructing the feature and label representation, respectively, and $\lambda_{1,2,3}$ control the relative importance of the regularization terms, which penalize large values for the Frobenius norm of the transformation matrices. Unfortunately, solve this problem directly is infeasible due to the function $\phi(\cdot)$ performs a mapping to a highly-dimensional space or even to an infinite-dimensional space, but, the first term of the loss function can be rewritten in terms of kernel matrices and employ the kernel trick [6].

$$\begin{aligned}\frac{\alpha}{2}\left\|\Phi\left(x_i\right)-\Phi\left(B\right)\tilde{W}_x W_x K\left(B,x_i\right)\right\|_F^2 = \frac{\alpha}{2}\Big(K\left(x_i,x_i\right)-2K\left(x_i,B\right)\tilde{W}_x W_x K\left(B,x_i\right)+ \\ K\left(B,x_i\right)^T W_x^T\tilde{W}_x^T K\left(B,B\right)\tilde{W}_x W_x K\left(B,x_i\right)\Big)\end{aligned}$$

Finally, we can make a change of variables a redefine the first term of the loss function ($\frac{\alpha}{2}\left(1-2z_i^T\tilde{z}_i+\tilde{z}_i^T K(B,B)\tilde{z}_i\right)$) and the structure (Fig. 1b), so that can be easily implemented in some deep learning framework.

Prediction. Once the parameters have been learned (coding and decoding matrices), we can use the model to predict the label representation $\tilde{y}$ from the feature representation x of a new unlabeled document by forward propagating it through the network.

Implementation details. The proposed method was implemented in the Keras [5] Framework, a high-level neural networks API written in Python and capable of running on top of either TensorFlow or Theano [10] libraries. The optimization is performed by stochastic gradient descent (SGD) with the RMSProp optimizer.

4 Experiments and Results

In this section, the proposed SS-OKSE algorithm will be evaluated in a multi-label annotation task under a semi-supervised setup. In order to compare our

algorithm, we use the same experimental setup proposed by Soleimani and Miller [9], where the performance of our method is compared against several supervised (PLLDA, miSVM and MISVM) and semi-supervised (ssLDA, and MLTM) methods in two different datasets of text documents:

Delicious. This dataset is composed of tagged web pages from the social bookmarking site delicious.com [13]. We adopt the subset proposed in [9] where the top 20 common tags are used as a class labels, constructing a subset portioned in 8350 documents for training and 4000 documents for testing. The documents are represented in a bag-of-word representation composed by a codebook of 8500 unique words obtained after applying Porter stemming and stopword removal.

Ohsumed. This collection contains medical abstracts from the MeSH categories of the year 1991 [13]. The specific task in this dataset is to categorizing 23 cardiovascular diseases categories. It is composed by 11122 training and 5388 test documents. Almost half of the documents have more than one label.

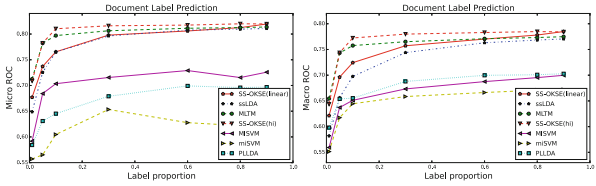
Experimental setup. For a fair comparison for all methods, where in some of them a suitable selection of the threshold is not trivial, the ROC AUC (Area Under the Curve) is used as the evaluation metric. Micro-ROC and Macro-ROC AUC are reported separately. In Micro-ROC, TPR and FPR are computed globally. In Macro-ROC, the ROC AUC is computed for each class across all documents and then the average is taken over all classes. While Micro-ROC may be dominated the bigger classes, Macro-ROC gives equal weight to all classes and better reveals performance on rare classes. For each dataset, a $(1-p)$ fraction is randomly selected from the training documents and their labels are removed. Then, for semi-supervised models, both labeled and unlabeled documents are used for training. But, for the purely supervised methods, only the remaining labeled documents are used for training. The annotation experiment is performed for different label proportions $p \in 0.01, 0.05, 0.1, 0.3, 0.6, 0.8, 0.9$ (five experiments are executed and the average is reported).

Determining the hyperparameters. Our model has 7 hyperparameters (α, β, λ_1, λ_2, λ_3, b, r). To properly determine the values of these hyperparameters, we randomly extract 20% of instances from the training set to validate the performance under a random (uniform) exploration in 30 different hyperparameter configurations trained with the remained 80% of training instances. The best configuration was chosen to be evaluated with the test partition. (this strategy have shown similar results than grid search while requires much fewer computation resources [3]).

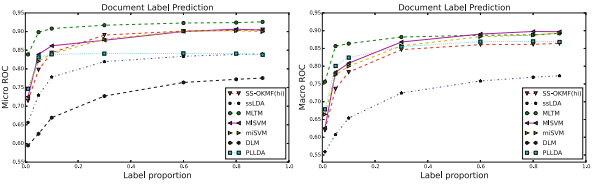
Multi-label annotation performance. Figure 2a presents the performance in ROC AUC in the Delicious dataset for different proportions of labeled documents. As we can see, the proposed SS-OKSE using a linear kernel presents competitive results against the other semi-supervised algorithms, and using a

histogram intersection (hi) kernel presents the best performance showing that a suitable kernel can help to model a good semantic space where non-labeled instances can be exploited. This becomes evident by achieving a performance very close to the maximum possible just with only 10% of labeled instances. Another important result is that this performance is obtained only with a budget composed by 500 instances randomly selected, this is only about the 6% of the training instances. Figure 2b presents the same previous experiment for the Delicious dataset. In Micro-ROC our method presents competitive results but the results in Macro-ROC suggest that the classes with the highest number of instances are dominating the learning. Still, it is important to emphasize that this result is obtained by using a budget of 2500 instances randomly selected (22% of the training instances) (Table 1).

Time-consuming and complexity. The main advantage of the SS-OKSE method is its highly scalable formulation. Table 2 shows the comparison between the proposed method and the state-of-the-art topic model method (we use the implementation supplied by the authors [9]). In both datasets SS-OKSE obtains remarkable speedups in comparison with MLTM under the same conditions (same computer with CPU device on a single core), but also thanks to its implementation, SS-OKSE is capable of running transparently in GPU devices achieving even more dramatic speedups, showing the ability of the proposed method to work on large scale datasets.



(a) Delicious dataset



(b) Ohsumed dataset

Fig. 2. Performance comparison in multi-label annotation.

Table 1. N: number of unique words, m: number of classes, l: number of instances, cardinality: average number of labels per instance.

			Training set		Test set	
Name	n	m	l	Cardinality	l	Cardinality
Delicious	8520	20	8251	2.89	3983	2.91
Ohsumed	13117	23	11122	1.65	5388	1.64

Table 2. Training time comparison (in minutes). For CPU device, only one core is using. Experiment running in a Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60 GHz with 16 cores, 64 GB RAM and an Nvidia Titan X.

Delicious					Ohsumed				
MLTM	SS-OKSE				MLTM	SS-OKSE			
CPU	CPU	Speedup	GPU	Speedup	CPU	CPU	Speedup	GPU	Speedup
961.3	43.5	22.1	5.1	186.9	1026	160.3	8.3	24.7	41.5

5 Conclusions

We have presented the SS-OKSE, a novel semi-supervised kernel semantic embedding method that uses a budget restriction to tackle the memory issue and computation time associated to kernel methods. The main advantage of using a budget in our method is that it allows us to save memory, since it is not necessary to store the complete kernel matrix, but a significantly smaller matrix defined by a budget keeping low computational requirements in large-scale problems. SS-OKSE is able to take advantage of annotated data to model a semantic low-dimensional space that preserves the separability of the original classes, and additionally, has the ability to exploit unlabeled instances for modeling the manifold structure of the data and use it to improve its performance in multi-label annotation tasks. The results confirm that the ability of the proposed method for modeling non-linearities can over-improve the performance in the multi-label annotation task.

Acknowledgment. Jorge A. Vanegas thanks for doctoral grant supports Colciencias 617/2013.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems, pp. 577–584 (2003)
2. Beltrán, V., Vanegas, J.A., González, F.A.: Semi-supervised dimensionality reduction via multimodal matrix factorization. In: Pardo, A., Kittler, J. (eds.) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. LNCS, vol. 9423, pp. 676–682. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25751-8_81

3. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13**, 281–305 (2012)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
5. Chollet, F.: Keras (2015). https://github.com/fchollet/keras
6. Chollet, F., et al.: Keras. GitHub (2015). https://github.com/keras-team/keras
7. Lee, H., Yoo, J., Choi, S.: Semi-supervised nonnegative matrix factorization. IEEE Sig. Process. Lett. **17**(1), 4–7 (2010)
8. Mcauliffe, J.D., Blei, D.M.: Supervised topic models. In: Advances in Neural Information Processing Systems, pp. 121–128 (2008)
9. Soleimani, H., Miller, D.J.: Semi-supervised multi-label topic models for document classification and sentence labeling. In: CIKM 2016, pp. 105–114. ACM (2016)
10. Theano Development Team. Theano: a Python framework for fast computation of mathematical expressions. arXiv e-prints, arXiv:1605.02688, May 2016
11. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 667–685. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-09823-4_34
12. Zhu, J., Ahmed, A., Xing, E.P.: Medlda maximum margin supervised topic models. J. Mach. Learn. Res. **13**(Aug), 2237–2278 (2012)
13. Zubiaga, A., García-Plaza, A.P., Fresno, V., Martínez, R.: Content-based clustering for tag cloud visualization. In: Social Network Analysis and Mining, ASONAM 2009, pp. 316–319. IEEE (2009)