Effective Semi-supervised Learning Based on Local Correlation

Xiao-Yu Zhang[1], Shupeng Wang[1*], Xin Jin[2*], Xiaobin Zhu[3*], and Binbin Li[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, China
[3] Beijing Technology and Business University, Beijing, China
[*] *Corresponding authors*
zhangxiaoyu@iie.ac.cn

Abstract. Traditionally, the manipulation of unlabeled instances is solely based on prediction of the existing model, which is vulnerable to ill-posed training set, especially when the labeled instances are limited or imbalanced. To address this issue, this paper investigate the local correlation based on the entire data distribution, which is leveraged as informative guidance to ameliorate the negative influence of biased model. To formulate the self-expressive property between instances within a limited vicinity, we develop the sparse self-expressive representation learning method based on column-wise sparse matrix optimization. Optimization algorithm is presented via alternating iteration. Then we further propose a novel framework, named semi-supervised learning based on local correlation, to effectively integrate the explicit prior knowledge and the implicit data distribution. In this way, the individual prediction from the learning model is refined by collective representation, and the pseudo-labeled instances are selected more effectively to augment the semi-supervised learning performance. Experimental results on multiple classification tasks indicate the effectiveness of the proposed algorithm.

Keywords: Semi-supervised Learning, Local Correlation, Self-expressive Representation, Sparse Matrix Optimization, Predictive Confidence.

1 Introduction

Machine learning has manifested its superiority in providing effective and efficient solutions for various applications [1][2][3][4][5]. In order to learn a robust model, the labeled instances are indispensable in that they convey precious prior knowledge and offer informative instruction. Unfortunately, manually labeling is labor-intensive and time-consuming. The cost associated with the labeling process often renders a fully labeled training set infeasible. In contrast, the acquisition of unlabeled instances is relatively inexpensive. One can easily have access to abundant unlabeled instances. As a result, when dealing with machine learning problems, we typically have to start with very limited labeled instances and plenty of unlabeled ones [6][7][8].

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_75

Recent researches indicate that the unlabeled instances, when used in conjunction with the labeled instances, can produce considerable improvement in learning performance. Semi-supervised learning [9][10], as a machine learning mechanism that jointly explores both labeled and unlabeled instances, has aroused widespread research attention. In semi-supervised learning, the exploration into unlabeled instances is largely dependent on the model trained with the labeled instances. On one hand, the insufficient or imbalanced labeled instances are inclined to lead to ill-posed learning model, which will consequently jeopardize the learning performance of semi-supervised learning. On the other hand, pair-wise feature similarity is widely used when estimating the data distribution, which is not necessarily plausible for semantic-level recommendation of class labels.

To address the aforementioned issues, this paper proposes a novel, named semi-supervised learning based on local correlation. Robust local correlation between the instances is estimated via sparse self-expressive representation learning, which formulates the self-expressive property between instances within a limited vicinity into a column-wise sparse matrix optimization problem. Based on the local correlation, an augmented semi-supervised learning framework is implemented, which takes into account both the explicit prior knowledge and the implicit data distribution. The learning sub-stages, including individual prediction, collective refinement, and dynamic model update with pseudo-labeling, are iterated until convergence. Finally, an effective learning model is obtained with encouraging experimental results on multiple classification applications.

2 Notation

In the text that follows, we let matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] = \boldsymbol{X}_L \cup \boldsymbol{X}_U \in \mathbb{R}^{m \times n}$ denote the entire dataset, where m is the dimension of features, and n is the total number of instances. In $\boldsymbol{X}$, each column $\boldsymbol{x}_i$ represents a m-dimensional instance. $\boldsymbol{X}_L \in \mathbb{R}^{m \times n_L}$ and $\boldsymbol{X}_U \in \mathbb{R}^{m \times n_U}$ are the labeled and unlabeled dataset, respectively, where n_L and n_U are the numbers of labeled and unlabeled instances, respectively. The corresponding class labels of the instances are denoted as matrix $\boldsymbol{Y} = \boldsymbol{Y}_L \cup \boldsymbol{Y}_U \in \mathbb{R}^{c \times n}$, where c is the number of classes, and $\boldsymbol{Y}_L \in \mathbb{R}^{c \times n_L}$ and $\boldsymbol{Y}_U \in \mathbb{R}^{c \times n_U}$ are the label matrices corresponding to the labeled and unlabeled dataset, respectively. For the labeled instance $\boldsymbol{x} \in \boldsymbol{X}_L$, its label $\boldsymbol{y} \in \boldsymbol{Y}_L$ is already known and denoted as a c-dimensional binary vector $\boldsymbol{y} = [y_1, \ldots, y_c]^T \in \{0,1\}^c$, whose ith element y_i $(1 \le i \le c)$ is a class indicator, i.e. $y_i = 1$ if instance $\boldsymbol{x}$ falls into class i, and $y_i = 0$ otherwise. For the unlabeled instance $\boldsymbol{x} \in \boldsymbol{X}_U$, its label $\boldsymbol{y} \in \boldsymbol{Y}_U$ is unrevealed and initially evaluated as $\boldsymbol{y} = \boldsymbol{0}$.

3 Robust Local Correlation Estimation

The locally linear property is widely applicable for smooth manifolds. In this scenario, an instance can be concisely represented by its close neighbors. As a result, the underlying local correlation is an effective reflection of the data distribution, and can be sub-

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_75

sequently leveraged as instructive guidance to improve the performance of model learning. Given the instance matrix $\boldsymbol{X}$, we develop a robust local correlation estimation method in a unsupervised fashion, which aims to infer a correlation matrix $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ based on the instances themselves regardless of their labels. The formulation is based on two major considerations. On one hand, an instance can be represented as a linear combination of the closely related neighbors. On the other hand, only a small number of neighbors are involved for the representation of an instance. In light of that, we estimate the robust local correlation between the instances via a novel Sparse Self-Expressive Representation Learning (SSERL), which is formulated as the follows.

$$\begin{aligned} &\min_{\boldsymbol{W}} \|\boldsymbol{W}^T\|_{2,1} \\ &\mathrm{s.t.} \boldsymbol{X} = \boldsymbol{XW}, \mathrm{diag}(\boldsymbol{W}) = \boldsymbol{0}, \boldsymbol{W} \geq 0 \end{aligned} \tag{1}$$

where the minimization of $\ell_{2,1}$-norm $\|\boldsymbol{W}^T\|_{2,1}$ ensures column sparsity of $\boldsymbol{W}$.

To make the problem more flexible, the equality constraint $\boldsymbol{X} = \boldsymbol{XW}$ is relaxed to allow expressive errors [11], and the corresponding objective is modified as follows.

$$\begin{aligned} &\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) = \|\boldsymbol{X} - \boldsymbol{XW}\|_F^2 + \lambda\|\boldsymbol{W}^T\|_{2,1} \\ &\mathrm{s.t.}\, \mathrm{diag}(\boldsymbol{W}) = \boldsymbol{0}, \boldsymbol{W} \geq 0 \end{aligned} \tag{2}$$

In $\mathcal{L}(\boldsymbol{W})$, the first term stands for the self-expressive loss and the second term is the column sparsity regularization. λ quantifies the tradeoff between the two terms.

The optimization problem (2) is not directly solvable. According to the general half-quadratic framework for regularized robust learning [12], we introduce an augmented cost function $\mathcal{A}(\boldsymbol{W}, \boldsymbol{p})$ as follows.

$$\mathcal{A}(\boldsymbol{W}, \boldsymbol{p}) = \|\boldsymbol{X} - \boldsymbol{XW}\|_F^2 + \lambda \mathrm{Tr}(\boldsymbol{WPW}^T) \tag{3}$$

where $\boldsymbol{p}$ is an auxiliary vector, and $\boldsymbol{P}$ is a diagonal matrix defined as $\boldsymbol{P} = \mathrm{diag}(\boldsymbol{p})$. The operator $\mathrm{diag}(\cdot)$ places a vector on the main diagonal of a square matrix.

With $\boldsymbol{W}$ given, the i-th entry of $\boldsymbol{p}$ is calculated as follows.

$$p_i = \frac{1}{2\|\boldsymbol{w}_i\|_2} \tag{4}$$

With $\boldsymbol{p}$ fixed, $\boldsymbol{W}$ can be optimized in a column-by-column manner as follows.

$$\boldsymbol{w}_i = (\boldsymbol{X}^T\boldsymbol{X} + \lambda p_i \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{x}_i \tag{5}$$

Based on (4) and (5), the auxiliary vector $\boldsymbol{p}$ and the correction matrix $\boldsymbol{W}$ are jointly optimized in an alternating iterative way. At the end of each iteration, the following post-processing is further implemented according to the constraints.

$$\begin{cases} \boldsymbol{W}_\Omega = 0, \Omega = \{(i,j) | 1 \leq i = j \leq n\} \\ \boldsymbol{W} = \max(\boldsymbol{W}, 0) \end{cases} \tag{6}$$

After convergence, the optimal correlation matrix $\boldsymbol{W}$ is obtained, which can serve as an informative clue for the revelation of the underlying data distribution and the construction of the subsequent model learning.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_75

4 Semi-supervised Learning Based on Local Correlation

Different from the traditional semi-supervised learning mechanism that solely depends on the classification model when exploring the unlabeled instances, the proposed SSL-LC takes into account both model prediction and local correlation. By iteration of the following three steps, SSL-LC is implemented in an effective way and the optimal learning model is obtained after convergence of the algorithm.

Step 1: individual label prediction by supervised learning.

As we know, for the labeled dataset $\boldsymbol{X}_L \in \mathbb{R}^{m \times n_L}$, the corresponding label set $\boldsymbol{Y}_L \in \mathbb{R}^{c \times n_L}$ is known beforehand. Using $(\boldsymbol{X}_L, \boldsymbol{Y}_L)$ as training dataset, the classification model $\mathcal{H}_\theta \colon \mathbb{R}^m \to \mathbb{R}^c$ can be obtained with off-the-shelf optimization methods. Specifically, probabilistic model can be applied based on the posterior distribution $P(\boldsymbol{y}|\boldsymbol{x};\theta)$ of label $\boldsymbol{y}$ conditioned on the input $\boldsymbol{x}$, where θ is the optimal parameter for $\mathcal{H}_\theta$ given $(\boldsymbol{X}_L, \boldsymbol{Y}_L)$. For the unlabeled instance $\boldsymbol{x} \in \boldsymbol{X}_U$, its label $\boldsymbol{y}$ is unknown and need to be predicted by the trained classification model $\mathcal{H}_\theta$. The prediction is given in the form of a c-dimensional vector $\widetilde{\boldsymbol{y}} = [P(y_1 = 1|\boldsymbol{x};\theta), \dots, P(y_c = 1|\boldsymbol{x};\theta)]^T \in [0,1]^c$. For the i-th entry $P(y_i = 1|\boldsymbol{x};\theta)$, larger value indicates higher probability that $\boldsymbol{x}$ falls into the i-th class with respect to $\mathcal{H}_\theta$, and vice versa. Based on the learning model $\mathcal{H}_\theta$, prediction can be made on each unlabeled instance individually. The predicted label set is collectively denoted as $\widetilde{\boldsymbol{Y}}_U$, which represents the classification estimation from the model point of view. With the dynamic update of model $\mathcal{H}_\theta$, the predicted label $\widetilde{\boldsymbol{Y}}_U$ is also dynamically renewed.

Step 2: collective label refinement by self-expressing.

In addition to the posterior probability estimated by model $\mathcal{H}_\theta$, the label of an unlabeled instance $\boldsymbol{x} \in \boldsymbol{X}_U$ can further be concisely represented by its closely related neighbors. The local correlation $\boldsymbol{W}$ calculated via SSERL reflects the underlying relevance between instances within the vicinity, and thus can serve as an informative guidance for self-expressive representation of labels. In this way, robust label refinement is achieved against potential classification errors. To be specific, the entire label matrix can be denoted as $\boldsymbol{Y}_p = \left[\boldsymbol{Y}_L, \widetilde{\boldsymbol{Y}}_U\right]$ after inference via classification. For further refinement, the local correlation matrix $\boldsymbol{W}$ is leveraged to obtain the self-expressive representation of labels in the form of $\boldsymbol{Y}_s = \boldsymbol{Y}_p\boldsymbol{W}$. By this means, the self-expressive property with respect to the instances is transferred to the labels, and the column-wise sparsity of $\boldsymbol{W}$ guarantees the concision of representation within a constrained vicinity. Then $\boldsymbol{Y}_s$ is normalized to obtain a legitimate probability estimation $\boldsymbol{Y}_n$, whose j-th column is calculated as:

$$[\boldsymbol{Y}_n]_j = \frac{[\boldsymbol{Y}_s]_j}{\max\limits_i [\boldsymbol{Y}_s]_{ij}} \tag{7}$$

Finally, since $\boldsymbol{Y}_L$ is already known and does not need to be estimated, the refined label matrix is calculated as:

$$\boldsymbol{Y}_r = [\boldsymbol{Y}_L, \boldsymbol{0}^{c\times n_U}] + \boldsymbol{Y}_n \odot [\boldsymbol{0}^{c\times n_L}, \boldsymbol{1}^{c\times n_U}] \tag{8}$$

where $\odot$ is the element-wise product of two matrices.

Step 3: semi-supervised model update by pseudo-labeling.

As discussed above, the effectiveness of semi-supervised learning stems from the comprehensive exploration on both labeled and unlabeled instances, in which the unlabeled instances with high predictive confidence are assigned with pseudo-labels and recommended to the learner as additional training data. The predictive confidence is the key measurement for selection of unlabeled instances. For the j-th instance, its predictive confidence is conveniently calculated as:

$$c_j = \max_i \left[\boldsymbol{Y}_n \odot [\boldsymbol{0}^{c\times n_L}, \boldsymbol{1}^{c\times n_U}]\right]_{ij} \tag{9}$$

which naturally filters out the labeled instances. Since $\boldsymbol{Y}_n$ is dependent on $\boldsymbol{Y}_p$ and $\boldsymbol{W}$, both individual classification prediction and collective local correlation are effectively integrated in the semi-supervised learning strategy. Based on the predictive confidence defined in (9), reliable and informative unlabeled instances can be selected and recommended for model update. The pseudo-label $\widehat{\boldsymbol{y}}_j$ associated with the j-th instance is defined as:

$$(\widehat{\boldsymbol{y}}_j)^i = \begin{cases} 1, \ i = \arg\max_i \left[\boldsymbol{Y}_n \odot [\boldsymbol{0}^{c\times n_L}, \boldsymbol{1}^{c\times n_U}]\right]_{ij} \\ 0, \ i \neq \arg\max_i \left[\boldsymbol{Y}_n \odot [\boldsymbol{0}^{c\times n_L}, \boldsymbol{1}^{c\times n_U}]\right]_{ij} \end{cases} \tag{10}$$

Using the pseudo-labeled instances as additional training data, the learning model $\mathcal{H}_\theta$ is re-trained, which brings about updated $\boldsymbol{Y}_p$ and $\boldsymbol{Y}_r$.

5 Experiments

To validate the effectiveness of SSL-LC, we apply it to classification tasks on malware [7] and patent [6] dataset respectively, in comparison with the following methods:

- Supervised learning (SL), which trains classifier based on the labeled dataset $\boldsymbol{T} = (\boldsymbol{X}_L, \boldsymbol{Y}_L)$, and arrives at individual prediction $\boldsymbol{Y}_p$ accordingly..
- Supervised learning with local correlation (SL-LC), which further refines the prediction with $\boldsymbol{W}$ and obtains $\boldsymbol{Y}_r$.
- Semi-supervised learning (SSL), which selects pseudo-labeled instances $\boldsymbol{R}$ based on unrefined prediction $\boldsymbol{Y}_p$, and updates classifier based on $\boldsymbol{T} \cup \boldsymbol{R}$.

Experiment 1: Comparison of Different Number of Labeled Instances. Firstly, we compare the classification performance with different number of labeled instances, i.e. $|\boldsymbol{T}|$. The classification performance is illustrated in Fig. 1 (1).

Detailed analysis of experimental results are as follows, where “>” stands for “outperform(s)”.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_75

- SL-LC > SL, SSL-LC > SSL. Under the instructive guidance of local correlation, the individual prediction on a single instance can be further refined via collective representation. Therefore, the classification results are more coherent to the intrinsic data distribution and less vulnerable to overfitting with local correlation refinement.
- SSL > SL, SSL-LC > SL-LC. Compared with supervised learning, semi-supervised learning further leverages the unlabeled instances to extend the training dataset, and thus receives higher classification performance.
- When the number of labeled instances is large enough, the difference between the four methods is negligible. It is indicated that the proposed SSL-LC is especially helpful for classification problems with insufficient labeled instances.

Experiment 2: Comparison of Different Number of Recommended Instances. We further compare the classification performance with different number of recommended instances, i.e. K, where SL and SL-LC are treated as special cases of SSL and SSL-LC with $K = 0$. The classification performance is illustrated in Fig. 1 (2).

As we can see, at first, the classification accuracy improves with the increase of K, because the model can learn from more and more instances. However, when K is large enough, further increase will lead to deterioration of classification performance. This results from the incorporation of the less confident pseudo-labeled instances, which inevitably brings about unreliable model.

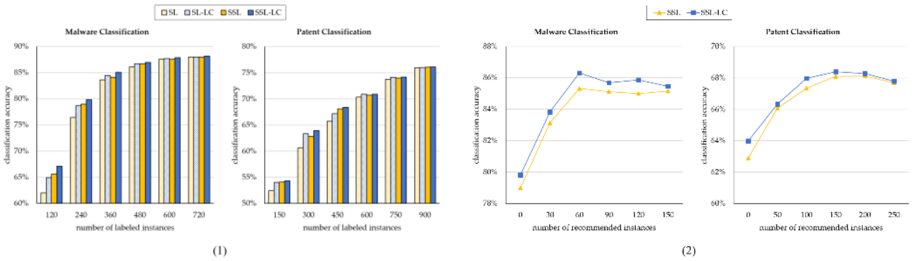

Fig. 1. The classification performance with different number of (1) labeled instances and (2) recommended instances.

6 Conclusion

In this paper, we have proposed an effective semi-supervised learning framework based on local correlation. Compared with traditional semi-supervised learning methods, the contributions of the work are as follows. Firstly, both the explicit prior knowledge and the implicit data distribution are integrated into a unified learning procedure, where the individual prediction from the dynamically updated learning model is refined by collective representation. Secondly, robust local correlation, rather than pair-wise similarity, is leveraged for model augment, which is formulated as a column-wise sparse matrix optimization problem. Last but not least, effective optimization is designed, in which the optimal solution is progressively reached in an iterative fashion. Experiments on multiple classification tasks indicate the effectiveness of the proposed algorithm.

7 Acknowledgement

This work was supported by National Natural Science Foundation of China (Grant 61501457, 61602517), Open Project Program of National Laboratory of Pattern Recognition (Grant 201800018), and Open Project Program of the State Key Laboratory of Mathematical Engineering and Advanced Computing (Grant 2017A05), and National Key Research and Development Program of China (Grant 2016YFB0801305).

References

1. Christopher, M.B., 2016. *Pattern Recognition and Machine Learning*. Springer-Verlag New York.
2. Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
3. Zhang, X., Xu, C., Cheng, J., Lu, H. and Ma, S., 2009. Effective annotation and search for video blogs with integration of context and content analysis. *IEEE Transactions on Multimedia*, *11*(2), pp.272-285.
4. Liu, Y., Zhang, X., Zhu, X., Guan, Q. and Zhao, X., 2017. Listnet-based object proposals ranking. *Neurocomputing*, *267*, pp.182-194.
5. Zhang, K., Yun, X., Zhang, X.Y., Zhu, X., Li, C. and Wang, S., 2016. Weighted hierarchical geographic information description model for social relation estimation. *Neurocomputing*, *216*, pp.554-560.
6. Zhang, X.Y., Wang, S. and Yun, X., 2015. Bidirectional active learning: a two-way exploration into unlabeled and labeled data set. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(12), pp.3034-3044.
7. Zhang, X.Y., Wang, S., Zhu, X., Yun, X., Wu, G. and Wang, Y., 2015. Update vs. upgrade: Modeling with indeterminate multi-class active learning. *Neurocomputing*, *162*, pp.163-170.
8. Zhang, X., 2014. Interactive patent classification based on multi-classifier fusion and active learning. *Neurocomputing*, *127*, pp.200-205.
9. Zhu, X. and Goldberg, A.B., 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, *3*(1), pp.1-130.
10. Soares, R.G., Chen, H. and Yao, X., 2012. Semisupervised classification with cluster regularization. *IEEE Trans. on Neural Networks and Learning Systems*, *23*(11), pp.1779-1792.
11. Zhang, X.Y., 2016. Simultaneous optimization for robust correlation estimation in partially observed social network. *Neurocomputing*, *205*, pp.455-462.
12. He, R., Zheng, W.S., Tan, T. and Sun, Z., 2014. Half-quadratic-based iterative minimization for robust sparse representation. *IEEE Transactions on PAMI*, *36*(2), 261-275.