Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Robust Semi-Supervised Learning through Label Aggregation

Yan Yan, Zhongwen Xu, Ivor W. Tsang, Guodong Long, Yi Yang
Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Australia
{yan.yan-3, zhongwen.xu}@student.uts.edu.au,{ivor.tsang, guodong.long, yi.yang}@uts.edu.au

Abstract

Semi-supervised learning is proposed to exploit both labeled and unlabeled data. However, as the scale of data in real world applications increases significantly, conventional semi-supervised algorithms usually lead to massive computational cost and cannot be applied to large scale datasets. In addition, label noise is usually present in the practical applications due to human annotation, which very likely results in remarkable degeneration of performance in semi-supervised methods. To address these two challenges, in this paper, we propose an efficient RObust Semi-Supervised Ensemble Learning (ROSSEL) method, which generates pseudo-labels for unlabeled data using a set of weak annotators, and combines them to approximate the ground-truth labels to assist semi-supervised learning. We formulate the weighted combination process as a multiple label kernel learning (MLKL) problem which can be solved efficiently. Compared with other semi-supervised learning algorithms, the proposed method has linear time complexity. Extensive experiments on five benchmark datasets demonstrate the superior effectiveness, efficiency and robustness of the proposed algorithm.

Introduction

Massive data can be easily collected from social networks and online services due to the explosion of Internet development. However, the vast majority of collected data are usually unlabeled and unstructured. Labeling a large amount of unlabeled data can be expensive. Therefore, it is natural to consider exploiting the abundance of unlabeled data to further improve the performance of algorithms. This has led to a rising demand for semi-supervised learning methods that leverage both labeled data and unlabeled data (Zhang et al. 2015; Lu et al. 2015; Zhu 2005; Chapelle, Schölkopf, and Zien 2006).

Semi-supervised learning (SSL) is an active research area and a variety of SSL algorithms have been proposed (Bennett, Demiriz, and others 1999; Blum and Chawla 2001; Chapelle, Weston, and Schölkopf 2002; Smola and Kondor 2003; Belkin, Matveeva, and Niyogi 2004; Li, Kwok, and Zhou 2010; Wang, Nie, and Huang 2014). However, many existing algorithms are faced with the scalability issue owing to the high complexity. For example, the complexity of LapSVM (Belkin, Niyogi, and Sindhwani 2006) is $O(n^3)$ due to the requirement for the inverse of a dense Gram matrix. TSVM in (Joachims 1999) treats the SVM problem as a sub-problem and infers the labels of unlabeled data via a label switch procedure, which may lead to a large number of iterations.

In addition to the scalability issue, SSL algorithms may suffer from label noise, leading to unreliable performance. In the SSL setting, there are usually only small amount of labeled data and a large proportion of unlabeled data. Even small mistakes in the human (non-expert) annotation process are likely to result in label noise. Thus robustness is particularly critical for SSL methods in many applications (Lu et al. 2015; Jing et al. 2015).

This paper focuses on the two aforementioned challenges of SSL, i.e. scalability and robustness. Inspired by crowdsourcing (Sheng, Provost, and Ipeirotis 2008; Snow et al. 2008), we propose an efficient RObust Semi-Supervised Ensemble Learning (ROSSEL) method to approximate ground-truth labels of unlabeled data through aggregating a number of pseudo-labels generated by low-cost *weak annotators*, such as linear SVM classifiers. Meanwhile, based on the aggregated labels, ROSSEL learns an inductive SSL classifier by Multiple Label Kernel Learning (MLKL) (Li et al. 2009). Unlike most existing SSL algorithms, the proposed ROSSEL requires neither expensive graph Laplacian nor iterative label switching. Instead, it only needs *one* iteration for label aggregation and can be solved by an SVM solver very efficiently. The major contributions are listed as follows,

- Leveraging an ensemble of low-cost supervised weak annotators, we propose ROSSEL to efficiently obtain a weighted combination of pseudo-labels of unlabeled data to approximate ground-truth labels to assist semi-supervised learning.
- Instead of simple label aggregation strategies used in crowdsourcing (e.g. majority voting), ROSSEL performs a weighted label aggregation using MLKL. Meanwhile it learns an inductive SSL classifier, which only requires *one* iteration and linear time complexity w.r.t. number of data and features.
- Complexity analysis of several competing SSL methods and the proposed method is provided.

Related Work

As large scale data are easily accessible, it is usually difficult to obtain sufficient supervision in practice. For instance, a feature selection algorithm is proposed in (Han et al. 2015) for video recognition where the number of labeled videos are limited. In (Gan et al. 2015), an action recognition method is proposed which does not exploit any positive examplars. The authors in (Li et al. 2013) propose a method to deal with weak-label learning tasks. In this paper, we focus on SSL problems.

Among SSL algorithms, graph-based methods are commonly used (Chapelle, Schölkopf, and Zien 2006). Many graph-based algorithms introduce the manifold structure by leveraging manifold regularization (Zhu, Ghahramani, and Lafferty 2003; Zhou et al. 2004; Belkin, Niyogi, and Sindhwani 2005; Sindhwani et al. 2005; Belkin, Niyogi, and Sindhwani 2006; Tsang and Kwok 2006; Sindhwani, Chu, and Keerthi 2007; Xu et al. 2010; Zhang et al. 2015). However, the complexity of building graph Laplacian is at least $O(n^2)$. Consequently, these graph-based algorithms are usually difficult to handle large scale datasets. Recently, the authors in (Wang, Nie, and Huang 2014) propose an adaptive SSL to optimize the weight matrix of the model and the label matrix simultaneously, which avoids expensive graph construction. There are some SSL methods exploiting pseudo-labels of unlabeled data. For instance, in (Lee 2013), pseudo-labels are used to make deep neural networks able to handle unlabeled data. The authors in (Bachman, Alsharif, and Precup 2014) propose to exploit pseudo-ensembles to produce models that are robust to perturbation. In (Deng et al. 2013), pseudo-labels are exploited in an image reranking framework regularized by multiple graphs. The authors in (Chang et al. 2014) formulate multi-label semi-supervised feature selection as a convex problem and propose an efficient optimization algorithm. A semi-supervised ranking and relevance feedback framework is proposed for multimedia retrieval in (Yang et al. 2012). In (Li, Kwok, and Zhou 2009), the authors propose a SVM-based SSL algorithm by exploiting the label mean. A cost-sensitive semi-supervised SVM is proposed in (Li, Kwok, and Zhou 2010). Although these methods avoid expensive graph Laplacian, they still require a number of iterations for training.

Ensemble learning is a supervised learning paradigm that trains a variety of learners on a given the training set, and derives a prediction from the votes of all its learners (Dietterich 2000). There are a number of most commonly used ensemble algorithms, including bagging (Breiman 1996), random forests (Breiman 2001) and boosting (Schapire and Freund 2012). Bagging is one of the most commonly used ensemble algorithms, where a number of bootstrap replicates are generated on the training set by bootstrap sampling, and a learner is trained on each bootstrap replicate. Ensemble learning methods can only handle labeled data.

The Proposed Model

Inspired by crowdsourcing methods (Sheng, Provost, and Ipeirotis 2008; Snow et al. 2008), we propose a new SSL algorithm that efficiently learns a classifier by leveraging both labeled and unlabeled data. Our proposed method consists of the two steps, namely label generation and label aggregation, illustrated in Figure 1. In the first stage, a set of weak annotators are trained and applied to unlabeled data to generate a set of pseudo-labels. In the second stage we combine the pseudo-labels to approximate the optimal labels of unlabeled data. In the meantime, weight vectors is derived, which enables ROSSEL to handle unseen data.

Label Generation

Low-cost, less-than-expert labels are easy to obtain from weak annotators in crowdsourcing (Sheng, Provost, and Ipeirotis 2008). Following the crowdsourcing framework, ROSSEL firstly generates a set of pseudo-labels for unlabeled data using ensemble learning. In this paper we focus on bagging to generate pseudo-labels.

Bagging is a simple and effective supervised ensemble learning algorithm, which produces a number of bootstrap replicates using bootstrap sampling. A weak learner is trained on each bootstrap replicate. By applying these weak learners on unlabeled data, a set of pseudo-labels can be derived. Bagging finally aggregates all the pseudo-labels by majority voting to generate predictions.

ROSSEL trains weak annotators using bootstrap sampling. Similar to crowdsourcing, we apply weak annotators on unlabeled data and obtain the resultant less-than-expert labels. The label generation procedure is illustrated in Figure 1.

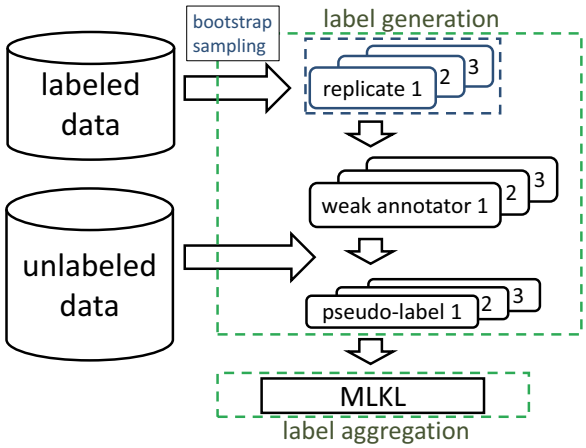


Figure 1: Illustration of the proposed ROSSEL.

Label Aggregation by MLKL

Considering a binary supervised learning scenario, let $\mathcal{D}_L = \{\mathbf{x}_i, y_i\}_{i=1}^{l}$ denotes the labeled set, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$ denotes the feature vector and the label of the i-th sample, respectively. A general objective function is formulated as follows

$$\min_{\mathbf{w}} \Omega(\mathbf{w}) + C\ell(\mathbf{w}), \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector, $\Omega(\mathbf{w})$ is the regularization term, $\ell(\mathbf{w})$ is a loss function and C is the regularization parameter. We focus on the ℓ_2-regularized hinge loss.

The objective function of hinge loss then can be specifically written as

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{l}\xi_i \quad s.t. \quad y_i\mathbf{w}^\top\mathbf{x}_i \geq 1-\xi_i, \xi_i \geq 0, i=1,...,l, \tag{2}$$

where ξ_i is the slack variable of the i-th instance.

SSL is aimed to exploit the abundant unlabeled data. Hence let $\mathcal{D}_U = \{\mathbf{x}_i\}_{i=l+1}^n$ denote the unlabeled set and we incorporate the information of unlabeled data into the objective function, which can be written as,

$$\min_{\tilde{\mathbf{y}}\in\mathcal{Y}}\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||_2^2 + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{i=l+1}^{n}\xi_i \quad s.t. \quad \tilde{y}_i\mathbf{w}^\top\mathbf{x}_i \geq 1-\xi_i,\ \xi_i \geq 0, i=1,...,n, \tag{3}$$

where C_1 and C_2 are the regularization parameters that control the tradeoff between model complexity, the cost generated by the labeled data, and the cost generated by the unlabeled data, and $\mathcal{Y} = \{\tilde{\mathbf{y}}|\tilde{\mathbf{y}} = [\mathbf{y}_L;\tilde{\mathbf{y}}_U], \tilde{\mathbf{y}}_U \in \{-1,+1\}^{n-l}\}$, where $\mathbf{y}_L \in \mathbb{R}^l$ represents the ground-truth label vector of labeled data, and $\tilde{\mathbf{y}}_U$ represents any possible labels of unlabeled data. Thus there are exponential possible values for $\mathbf{y}_U$, i.e. the labels of unlabeled data, which is intractable to directly optimize.

By introducing dual variables $\boldsymbol{\alpha} \in \mathbb{R}^n$, the Lagrangian of Equation (3) can be obtained by

$$\mathcal{L}(\mathbf{w},\boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{w}||_2^2 + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{i=l+1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1-\xi_i-\tilde{y}_i\mathbf{w}^\top\mathbf{x}_i). \tag{4}$$

By setting the derivatives of $\mathcal{L}$ w.r.t. $\mathbf{w}$ and ξ_i as 0, the Lagrangian can be updated as below,

$$\mathcal{L} = -\frac{1}{2}\boldsymbol{\alpha}^\top\Big((XX^\top)\odot\tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top\Big)\boldsymbol{\alpha} + \mathbf{1}^\top\boldsymbol{\alpha}, \tag{5}$$

where $\boldsymbol{\alpha} \in \mathcal{A}$ and $\mathcal{A} = \{\boldsymbol{\alpha}|0 \leq \alpha_i \leq C_1, 0 \leq \alpha_j \leq C_2, 1 \leq i \leq l, l+1 \leq j \leq n\}$. We can then replace the inner minimization problem of Problem (3) by its dual as below,

$$\min_{\tilde{\mathbf{y}}\in\mathcal{Y}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}^\top\Big((XX^\top)\odot\tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top\Big)\boldsymbol{\alpha} + \mathbf{1}^\top\boldsymbol{\alpha}, \tag{6}$$

where $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]^\top$. It is usually difficult to optimize $\tilde{\mathbf{y}}$ due to the significant number of possible values. Inspired by ideas from crowdsourcing, which obtain sufficiently qualified labels on unlabeled data by exploiting a set of weak annotators, we propose to solve Problem (6) by MLKL (Li et al. 2013; 2009).

Definition 1. *Given a size-M label set $\{\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, ..., \tilde{\mathbf{y}}_M\}$, multiple label kernel learning (MLKL) refers to the problem as below,*

$$\min_{\boldsymbol{\mu}\in\mathcal{U}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}^\top\Big((XX^\top)\odot(\sum_{m=1}^{M}\mu_m\tilde{\mathbf{y}}_m\tilde{\mathbf{y}}_m^\top)\Big)\boldsymbol{\alpha} + \mathbf{1}^\top\boldsymbol{\alpha}, \tag{7}$$

which aims to find a weighted combination of the label kernels $\sum_{m=1}^{M}\mu_m\tilde{\mathbf{y}}_m\tilde{\mathbf{y}}_m^\top$ to approximate the ground-truth label kernel $\tilde{\mathbf{y}}^*\tilde{\mathbf{y}}^{*\top}$, where $\mathcal{U} = \{\boldsymbol{\mu}|\sum_{m=1}^{M}\mu_m = 1,\ \mu_m \geq 0\}$, $\mathcal{A} = \{\boldsymbol{\alpha}|0 \leq \alpha_i \leq C_1, 0 \leq \alpha_j \leq C_2, 1 \leq i \leq l, l+1 \leq j \leq n\}$, and $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_M]^\top$ denotes the weight vector of base label kernels.

Similar to crowdsourcing, a set of pseudo-labels of unlabeled data are generated in the first step by bootstrap sampling. In the second step, we propose to obtain the SSL classifier by MLKL. Assume that there are M pseudo-labels, namely $\mathcal{Y}_M = \{\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, ..., \tilde{\mathbf{y}}_M\}$, then we can complete the primal formulation of Problem (7) as,

$$\min_{\boldsymbol{\mu}\in\mathcal{U},\mathbf{w}_m,\xi} \frac{1}{2}\sum_{m=1}^{M}\frac{1}{\mu_m}||\mathbf{w}_m||_2^2 + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{i=l+1}^{n}\xi_i \quad s.t. \quad \sum_{m=1}^{M}\tilde{y}_{mi}\mathbf{w}_m^\top\mathbf{x}_i \geq 1-\xi_i, \xi_i \geq 0, i=1,...,n, \tag{8}$$

where $\tilde{y}_{mi}$ denotes the label for the i-th sample in $\mathbf{y}_m$.

By setting $\hat{\mathbf{w}} = [\frac{\mathbf{w}_1}{\sqrt{\mu_1}}, ..., \frac{\mathbf{w}_M}{\sqrt{\mu_M}}]^\top$, $\hat{\mathbf{x}}_i = [\sqrt{\mu_1}\mathbf{x}_i, \sqrt{\mu_2}\tilde{y}_{1i}\tilde{y}_{2i}\mathbf{x}_i, ..., \sqrt{\mu_T}\tilde{y}_{1i}\tilde{y}_{Mi}\mathbf{x}_i]^\top$, and $\hat{\mathbf{y}} = \tilde{\mathbf{y}}_1$, the primal problem of MLKL (7) becomes

$$\min_{\hat{\mathbf{w}},\xi} \frac{1}{2}||\hat{\mathbf{w}}||_F^2 + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{i=l+1}^{n}\xi_i \quad s.t. \quad \hat{y}_i\hat{\mathbf{w}}^\top\hat{\mathbf{x}}_i \geq 1-\xi_i, \xi_i \geq 0, i=1,...,n. \tag{9}$$

Problem (9) is similar to the primal of a standard SVM problem, and can be easily solved by existing SVM packages, such as LIBLINEAR. Compared to Problem (3), Problem (9) can be solved very efficiently.

ROSSEL is easy to extend to cope with multiclass problems by applying the one-vs-all strategy. The detailed ROSSEL algorithm for a multiclass case can be found in Algorithm 1.

Complexity Analysis

There are two main stages in the proposed method, namely label generation and label aggregation. In the label generation step, M weak annotators are trained. Weak annotators can be any cheap learner. In our experiments, we use LIBLINEAR to train linear SVMs as the weak annotators. Hence, this leads to a complexity of $O(Mnd)$ where n and d stand for the number and the dimension of data respectively. In the label aggregation step, MLKL can be solved according to Problem (9) by LIBLINEAR (Fan et al. 2008), and $\boldsymbol{\mu}$, the coefficient of base label kernels, can be simply updated by closed-form solution, which results in the complexity of $O(Mnd)$. Compared with many other SSL methods that require a number iterations for label switching and model training, the proposed ROSSEL only requires *one* iteration. Therefore, the overall complexity of ROSSEL is $O(Mnd)$, which does not rely on T, the number iterations.

In Table 1, we list the complexity of various SSL algorithms, including LapSVM (Belkin, Niyogi, and Sindhwani

Table 1: Comparison of complexity of the proposed method and other related SSL methods.

Mthods	LapSVM	LapRLS	meanS3VM	CS4VM	ASL	ROSSEL
Complexity	$O(n^3d)$	$O(n^3d)$	$O(n^2dT)$	$O(n^2dT)$	$O(nd^2T)$	$O(Mnd)$

In this table, n, d, M and T represent the number of data, the dimension of data, the number of weak annotators and the number of iterations of the algorithm respectively.

Algorithm 1 RObust Semi-Supervised Ensemble Learning (ROSSEL)

1: Initialize M, the number of weak annotators.
2: **for** $k = 1$ to K **do**
3: Sample M bootstrap replicates $\{(\bar{\mathbf{X}}_1, \bar{\mathbf{y}}_{k1}), (\bar{\mathbf{X}}_2, \bar{\mathbf{y}}_{k2}), ..., (\bar{\mathbf{X}}_T, \bar{\mathbf{y}}_{kM})\}$ from the labeled set $\mathcal{D}_L$.
4: **for** $m = 1$ to M **do**
5: Train an SVM model $\mathcal{M}_{km}$ on $\bar{X}_m$ and $\bar{\mathbf{y}}_{km}$.
6: Derive $\tilde{\mathbf{y}}_{km}$ by predicting on the unlabeled data X_U using $\mathcal{M}_{km}$.
7: Add $\tilde{\mathbf{y}}_{km}$ into the working set $\mathcal{Y}_{kM}$
8: **end for**
9: Compute $\{\mathbf{w}_{k1}, \mathbf{w}_{k2}, ..., \mathbf{w}_{kM}\}$ and $\boldsymbol{\mu}_k$ by solving Problem (8).
10: Calculate prediction $p_{jk} = \sum_{m=1}^{M} \mu_{km} \mathbf{w}_{km}^\top \mathbf{x}_j$ for a test data $\mathbf{x}_j$.
11: **end for**
12: Choose the class label for $\mathbf{x}_j$ by $\arg\max_k \{p_{jk}\}_{k=1}^K$.

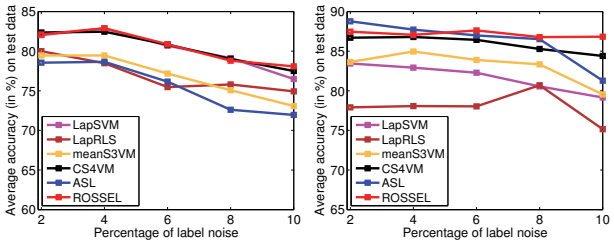
2006), LapRLS (Belkin, Niyogi, and Sindhwani 2006), meanS3VM (Li, Kwok, and Zhou 2009), CS3VM (Li, Kwok, and Zhou 2010) and ASL (Wang, Nie, and Huang 2014). LapSVM and LapRLS have high complexity w.r.t. the number of instances n due to the inverse of a dense Gram matrix. Note that meanS3VM, CS4VM and ASL require to update their models iteratively. Consequently, their complexity contains T. It can be expensive if a large number of iterations is required.

Experiments

In this section, we demonstrate the robustness and performance of the proposed algorithm by comparing with eight baselines. These baselines include three supervised learning methods, namely LIBLINEAR (Fan et al. 2008), LIBSVM (Chang and Lin 2011), ensemble LIBSVM, and five SSL algorithms, namely LapSVM (Belkin, Niyogi, and Sindhwani 2006), LapRLS (Belkin, Niyogi, and Sindhwani 2006), meanS3VM (Li, Kwok, and Zhou 2009), CS4VM (Li, Kwok, and Zhou 2010) and ASL (Wang, Nie, and Huang 2014). In total six datasets are used, namely CNAE9, dna, connect4, protein and rcv1-train and rcv1-all. Three experiments are performed, which respectively investigate the resistance to label noise, performance on various scale datasets and the impact of different numbers of weak annotators in ROSSEL. All experiments are conducted on a workstation with an Intel(R) CPU (Xeon(R) E5-2687W v2 @ 3.40GHz) and 32 GB memory.

Table 2: Data statistics.

Datasets	# train	# test	# features	# classes
CNAE9	800	280	856	9
dna	2,559	627	180	3
connect4-10k	8,000	2,000	126	3
protein	19,200	5,187	357	3
rcv1-train	12,384	3,114	47,236	38
rcv1-all	420,000	111,920	47,236	40



(a) CNAE9 (b) dna

Figure 2: Average accuracy on the CNAE9 and dna datasts over 10 runs when label noise is present.

Datasets

Five UCI datasets including CNAE9, dna, connect4, protein and rcv1 are used in the experiments. Among them, CNAE9 and dna are two small scale datasets that every competing method is able to handle. Protein, connect4 and rcv1 are large scale datasets which are used to investigate both the accuracy and scalability of competing methods. The size of connect4 and rcv1, which contain 67,557 and 534,135 samples respectively, is very large for SSL algorithms. Consequently, for the convenience of comparison on connect4, we generate a new dataset called connect4-10k by sampling 10,000 instances from connect4 at random. We report results of the rcv1 dataset on both the standard training set and the full set.

In all experiments, to simulate the SSL scenario, we randomly sample three disjointed subsets from each dataset as the labeled set (5% samples), unlabeled set (75% samples) and test set (20%). More information about the six datasets is listed in Table 2. We report accuracy as the evaluation metric for comparison in all tables and figures.

Resistance to Label Noise

In this experiment, we investigate the resistance of SSL algorithms to label noise on the CNAE9 and dna datasets. We randomly select 2%, 4%, ..., 10% labels from the labeled set

Table 3: Average accuracy (±Standard Deviation(%)) over 10 runs.

Methods	CNAE9	dna	connect4-10k	protein	rcv1-train	rcv1-all
LIBLINEAR	82.86(±2.56)	84.78(±1.52)	64.40(±1.73)	60.54(±0.64)	74.45(±1.89)	87.56(±0.09)
LIBSVM	83.04(±2.94)	85.96(±1.42)	63.43(±2.43)	61.84(±1.31)	74.93(±1.88)	87.57(±0.12)
ensemble-10SVM	79.75(±2.41)	83.32(±1.38)	65.26(±2.54)	60.78(±1.40)	72.39(±1.53)	87.46(±0.09)
ensemble-50SVM	81.56(±2.42)	84.63(±1.84)	65.70(±1.99)	60.91(±0.85)	73.17(±1.90)	87.60(±0.08)
LapSVM	85.33(±3.13)	85.63(±1.28)	64.39(±1.82)	60.46(±0.85)	74.91(±1.90)	*
LapRLS	**85.47(±2.72)**	85.84(±1.23)	63.41(±1.63)	60.72(±0.61)	74.55(±1.92)	*
meanS3VM	83.12(±3.57)	85.04(±1.17)	–	–	–	–
CS4VM	84.93(±2.98)	88.04(±1.12)	62.04(±2.14)	–	–	–
ASL	82.61(±2.15)	**90.03(±0.98)**	60.83(±1.41)	58.94(±1.19)	*	*
ROSSEL10	85.11(±2.42)	88.50(±1.91)	67.89(±1.16)	61.88(±1.34)	**79.22(±2.00)**	**89.20(±0.15)**
ROSSEL50	85.04(±3.14)	88.52(±1.54)	**68.20(±0.98)**	**62.33(±0.90)**	78.77(±2.25)	89.18(±0.11)

We report the results of ensemble-SVM and ROSSEL with both 10 and 50 weak annotators. Semi-supervised methods with maximum accuracy are in **bold**. Some of the compared algorithms either require much memory (indicated by "*" in the above table) or very expensive in computation (e.g. more than a day, indicated by "–" in the above table). Therefore, these algorithms can not be applied to the large datasets such as the rcv1-all dataset.

Table 4: Average training time (in seconds) over 10 runs.

Methods	CNAE9	dna	connect4-10k	protein	rcv1-train	rcv1-all
LIBLINEAR	0.0008	0.0009	0.0909	0.0126	0.3405	1.6855
LIBSVM	0.0052	0.0385	0.1408	0.2387	1.3338	672.4409
ensemble-10SVM	0.0060	0.0136	0.0487	0.2329	2.1081	33.2070
ensemble-50SVM	0.0224	0.0405	0.3919	0.8019	14.5482	119.0243
LapSVM	**0.1596**	7.0668	14.3528	152.9257	494.4695	*
LapRLS	0.1715	7.0214	13.0248	152.8537	420.5253	*
meanS3VM	2.8588	13.8941	–	–	–	–
CS4VM	1.3219	9.5178	539.8876	–	–	–
ASL	3.4355	16.3261	115.6894	1748.2612	*	*
ROSSEL10	0.2123	**0.2271**	**0.7955**	**3.4457**	**45.5584**	**815.0660**
ROSSEL50	0.5481	1.4133	3.2811	16.5558	336.4487	6024.5965

We report the results of ensemble-SVM and ROSSEL with both 10 and 50 weak annotators. Semi-supervised methods with minimum training time are in **bold**. Some of the compared algorithms either require much memory (indicated by "*" in the above table) or very expensive in computation (e.g. more than a day, indicated by "–" in the above table). Therefore, these algorithms can not be applied to the large datasets such as the rcv1-all dataset.

and switch them to wrong labels as label noise. The resultant accuracy reported in Figure 2 demonstrates that our algorithm can be more resistant to label noise than other baselines used in the experiment.

Comparison of Accuracy and Scalability

In this experiment, we investigate the accuracy and scalability of SSL algorithms. We compare the proposed algorithm with eight other methods, including three supervised learning algorithms and five SSL methods. The three supervised learning baselines are listed as below:

- LIBLINEAR (Fan et al. 2008) is a supervised linear SVM baseline, efficient for large scale data. In the experiment, we tune two types of SVM including L2-regularized L2-loss and L2-regularized L1-loss and report the best results. We apply the one-vs-all strategy for all experiments.
- LIBSVM (Chang and Lin 2011) is a supervised non-linear SVM baseline, which is usually slower than LIBLINEAR when kernels are present. In the experiment, we tune various kernels, including the linear kernel, polynomial kernel, Gaussian kernel and sigmoid kernel. We apply the one-vs-all strategy for all experiments.
- Ensemble-SVM is an ensemble supervised learning baseline, by which we demonstrate the effectiveness of the proposed SSL method. Each of the base classifier is trained by LIBLINEAR on a bootstrap replicate. The predicted label on a test instance is computed by plurality voting of all base classifier.

The five SSL competing methods are listed as follows:

- LapSVM (Belkin, Niyogi, and Sindhwani 2006) is a graph-based SSL algorithm. The objective function of SVMs is regularized by graph Laplacian.
- LapRLS (Belkin, Niyogi, and Sindhwani 2006), similar to LapSVM, is regularized by graph Laplacian. The objective function is based on the least squared loss.
- meanS3VM (Li, Kwok, and Zhou 2009), instead of estimating the label of each unlabeled data, exploits the label means of unlabeled data, and maximizes the margin between the label means.
- CS4VM (Li, Kwok, and Zhou 2010) is a cost-sensitive semi-supervised SVM algorithm, which treats various

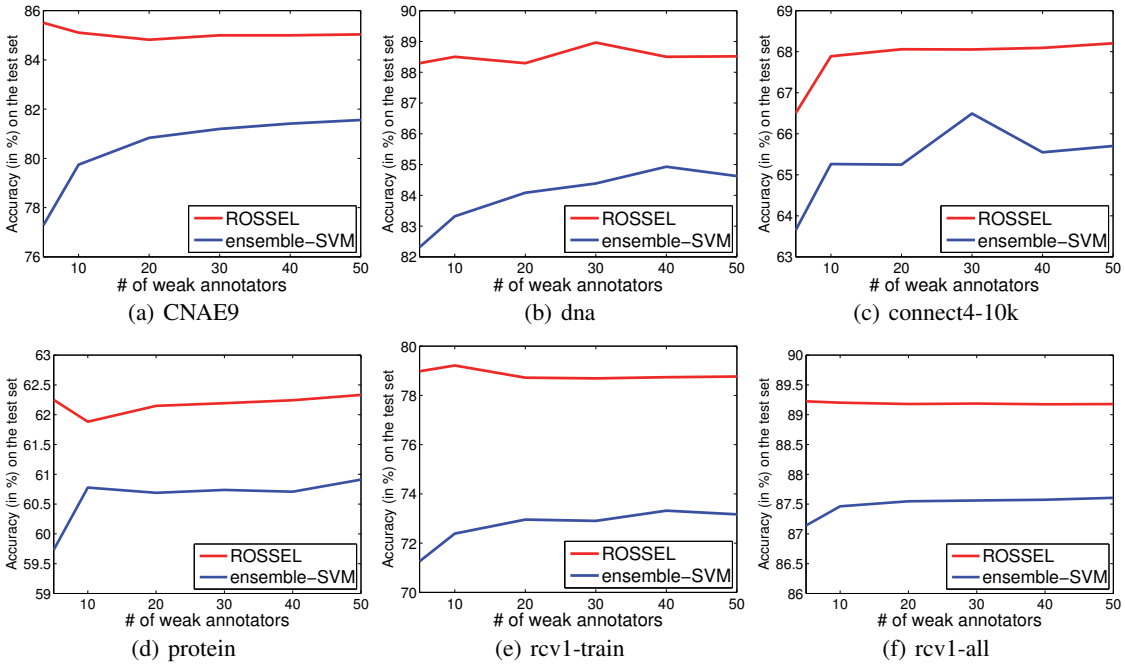


Figure 3: Average accuracy over 10 runs on various datasets with different number of weak annotators.

misclassification errors with different costs.

- ASL (Wang, Nie, and Huang 2014) is a recently proposed SSL method that avoids expensive graph construction and adaptively adjusts the weights of data, which can be robust to boundary points.

In this experiment, we use Gaussian kernel $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2})$ to compute the kernel matrix. The kernel parameter σ is fixed as 1, and all feature matrices are normalized before the experiment. We select from the range of $\{10^{-5}, 10^{-3}, 10^{-1}, 10^{0}, 10^{1}, 10^{3}, 10^{5}\}$ for the parameters to be tuned in all methods. We empirically set the parameter k-nearest neighbour as 5 for the graph-based methods, LapSVM and LapRLS.

Ensemble-SVM and ROSSEL are two ensemble based methods. In this experiment, we report the results of these two methods with both 10 and 50 weak annotators. When sampling, we bootstrap 50% labeled data into a bootstrap replicate.

For comparison, we perform the experiment 10 times on various splits of the labeled, unlabeled and test sets. Average accuracy on the test set and average training time of all competing methods over 10 runs are reported in Table 3 and Table 4 respectively. Results in the two tables demonstrate that the proposed method is very competitive in terms of accuracy and scalability. SSL algorithms usually suffer from poor scalability. As can be seen from the results, even on the full rcv1 dataset that contains more than 400,000 training examples with 47,236 features, ROSSEL provides promising accuracy within much less training time.

Impact of Number of Weak Annotators

In this experiment, we study the effect of various numbers of weak annotators used in the two ensemble based methods, ROSSEL and ensemble-SVM. We perform this experiment on all the six datasets.To investigate the influence of different numbers of weak annotators, 5, 10, 20, 30, 40, 50 weak annotators are used in these two methods. We run the experiment over 10 different splits of labeled, unlabeled and the test sets. The accuracy on test data of different numbers of weak annotators of the two algorithms is reported in Figure 3.

As observed, ensemble-SVM usually performs better with more weak annotators. However, our method with different numbers of weak annotators gives very close performance. This observation demonstrates that our algorithm is stable and will provide competitive performance even when there are a small number of weak annotators involved.

Conclusions

SSL is proposed to improve the performance by exploiting both labeled data and unlabeled data. It plays an increasingly crucial role in practical applications due to the rapid boosting of the volume of data. However, conventional SSL algorithms usually suffer from the poor efficiency and may degenerate remarkably when label noise is present. To address these two challenges, we propose ROSSEL to approximate ground-truth labels for unlabeled data though the weighted aggregation of pseudo-labels generated by low-cost weak annotators. Meanwhile ROSSEL trains an inductive SSL model. We formulate the label aggregation problem as a multiple label kernel learning (MLKL) problem which can be solved very efficiently. The complexity of ROSSEL

is much lower than related SSL methods. Extensive experiments are performed on five benchmark datasets to investigate the robustness, accuracy and efficiency of SSL methods.

Acknowledgement

This project is partially supported by the ARC Future Fellowship FT130100746, ARC grant LP150100671, the ARC DECRA project DE130101311 and the ARC discovery project DP150103008.

References

Bachman, P.; Alsharif, O.; and Precup, D. 2014. Learning with pseudo-ensembles. In *NIPS*.

Belkin, M.; Matveeva, I.; and Niyogi, P. 2004. Regularization and semi-supervised learning on large graphs. In *COLT*.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2005. On manifold regularization. In *AISTATS*.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR* 7:2399–2434.

Bennett, K.; Demiriz, A.; et al. 1999. Semi-supervised support vector machines. In *NIPS*.

Blum, A., and Chawla, S. 2001. Learning from labeled and unlabeled data using graph mincuts. In *ICML*.

Breiman, L. 1996. Bagging predictors. *Machine learning* 24(2):123–140.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1):5–32.

Chang, C.-C., and Lin, C.-J. 2011. Libsvm: A library for support vector machines. *TIST* 2(3):27.

Chang, X.; Nie, F.; Yang, Y.; and Huang, H. 2014. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*.

Chapelle, O.; Schölkopf, B.; and Zien, A. 2006. *Semi-supervised learning*. MIT press Cambridge.

Chapelle, O.; Weston, J.; and Schölkopf, B. 2002. Cluster kernels for semi-supervised learning. In *NIPS*.

Deng, C.; Ji, R.; Liu, W.; Tao, D.; and Gao, X. 2013. Visual reranking through weakly supervised multi-graph learning. In *ICCV*.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In *Multiple classifier systems*. Springer. 1–15.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. Liblinear: A library for large linear classification. *JMLR* 9:1871–1874.

Gan, C.; Lin, M.; Yang, Y.; Zhuang, Y.; and G Hauptmann, A. 2015. Exploring semantic inter-class relationships (sir) for zero-shot action recognition. In *AAAI*.

Han, Y.; Yang, Y.; Yan, Y.; Ma, Z.; Sebe, N.; and Zhou, X. 2015. Semisupervised feature selection via spline regression for video semantic recognition. *TNNLS* 26(2):252–264.

Jing, X.-Y.; Liu, Q.; Wu, F.; Xu, B.; Zhu, Y.; and Chen, S. 2015. Web page classification based on uncorrelated semi-supervised intra-view and inter-view manifold discriminant feature extraction. In *IJCAI*.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *ICML*.

Lee, D.-H. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*.

Li, Y.-F.; Tsang, I. W.; Kwok, J. T.; and Zhou, Z.-H. 2009. Tighter and convex maximum margin clustering. In *AISTATS*, 344–351.

Li, Y.-F.; Tsang, I. W.; Kwok, J. T.; and Zhou, Z.-H. 2013. Convex and scalable weakly labeled svms. *JMLR* 14(1):2151–2188.

Li, Y.-F.; Kwok, J. T.; and Zhou, Z.-H. 2009. Semi-supervised learning using label mean. In *ICML*.

Li, Y.-F.; Kwok, J. T.; and Zhou, Z.-H. 2010. Cost-sensitive semi-supervised support vector machine. In *AAAI*.

Lu, Z.; Gao, X.; Wang, L.; Wen, J.-R.; and Huang, S. 2015. Noise-robust semi-supervised learning by large-scale sparse coding. In *AAAI*.

Schapire, R. E., and Freund, Y. 2012. *Boosting: Foundations and Algorithms*. The MIT Press.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD*.

Sindhwani, V.; Niyogi, P.; Belkin, M.; and Keerthi, S. 2005. Linear manifold regularization for large scale semi-supervised learning. In *Proc. of the 22nd ICML Workshop on Learning with Partially Classified Training Data*.

Sindhwani, V.; Chu, W.; and Keerthi, S. S. 2007. Semi-supervised gaussian process classifiers. In *IJCAI*.

Smola, A. J., and Kondor, R. 2003. Kernels and regularization on graphs. In *Annual Conference on Computational Learning Theory*.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP*.

Tsang, I. W., and Kwok, J. T. 2006. Large-scale sparsified manifold regularization. In *NIPS*.

Wang, D.; Nie, F.; and Huang, H. 2014. Large-scale adaptive semi-supervised learning via unified inductive and transductive model. In *KDD*.

Xu, Z.; King, I.; Lyu, M. R.-T.; and Jin, R. 2010. Discriminative semi-supervised feature selection via manifold regularization. *TNN* 21(7):1033–1047.

Yang, Y.; Nie, F.; Xu, D.; Luo, J.; Zhuang, Y.; and Pan, Y. 2012. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *TPAMI* 34(4):723–742.

Zhang, K.; Lan, L.; Kwok, J.; Vucetic, S.; and Parvin, B. 2015. Scaling up graph-based semisupervised learning via prototype vector machines. *TNNLS* 26(3):444–457.

Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; and Schölkopf, B. 2004. Learning with local and global consistency. *NIPS* 16(16):321–328.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*.

Zhu, X. 2005. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.