Journal of the Korean Data & Information Science Society
2011, **22**(1), 99–105

한국데이터정보과학회지

Semi-supervised learning using similarity and dissimilarity[†]

Kyung Ha Seok[1]

[1]Department of Data Science, Institute of Statistical Information, Inje university

Received 22 October 2010, revised 15 December 2010, accepted 20 December 2010

Abstract

We propose a semi-supervised learning algorithm based on a form of regularization that incorporates similarity and dissimilarity penalty terms. Our approach uses a graph-based encoding of similarity and dissimilarity. We also present a model-selection method which employs cross-validation techniques to choose hyperparameters which affect the performance of the proposed method. Simulations using two types of data sets demonstrate that the proposed method is promising.

Keywords: Dissimilarity penalty term, generalized cross validation, kernel estimation, manifold regularization, semi-supervised learning.

1. Introduction

The promising empirical success of semi-supervised learning (SSL) algorithms in favorable situations has triggered several recent attempts (Lafferty and Wasserman, 2007; Niyogi, 2008) at developing a theoretical understanding of SSL. In a recent paper by Singh *et al*. (2008), it was established using a finite sample analysis that if the complexity of the distributions under consideration is too high to be learned using l labeled data points, but is small enough to be learned using $u(>> l)$ unlabeled data points, then semi-supervised learning can improve the performance of a supervised learning task.

There have also been many successful practical SSL algorithms as summarized in Chapelle *et al*. (2006), Zhu (2005) and Zhu and Goldberg (2009). These theoretical analysis and practical algorithms often assume that the data forms clusters or resides in a single manifold. Also, they implement similarity graphs only. Goldberg *et al.* (2007) proposed using both dissimilarity and similarity. They assume that the dissimilarity knowledge they use is known and noisy. In practical application, it is hard to meet these assumptions of the dissimilarity knowledge.

In this paper, we propose a new SSL which incorporates both similarity and dissimilarity. We start with graph-based SSL (Belkin *et al*., 2006; Zhu *et al*., 2003) which allows a natural

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0016813).

[1] Professor, Department of Data Science, Institute of Statistical Information Inje University, Gyungnam 621-749, Korea. E-mail: statskh@inje.ac.kr.

combination of similarity and dissimilarity. Existing graph-based SSL methods encode label similarity knowledge but they cannot handle dissimilarity easily, as we show in Section 2. We define a mixed graph to accommodate both in Section 3. In Section 4, we present a model selection method, and we present experimental results in Section 5.

2. Manifold regularization

Let there be n items, of which l are labeled : $\{(x_1, y_1), ..., (x_l, y_l), x_{l+1}, ..., x_n\}$, $x_i \in R^d$, $y_i \in \{-1, 1\}$. Existing graph-based SSL methods assume that a graph over the n item is given. The graph is represented by an $n \times n$ matrix W, where w_{ij} is the nonnegative edge weight between items i, j. Similar items have large weights, assuming that they have same labels. Such facts can be represented as a penalty term on the discriminant function $f : X \to R$:

$$\frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f(x_i) - f(x_j))^2. \tag{2.1}$$

Minimization of (2.1) tends to force $f(x_i) \approx f(x_j)$ when w_{ij} is large; therefore existing graph-based methods are able to encode label similarity. The penalty (2.1) can be written in quadratic form $\boldsymbol{f}'L\boldsymbol{f}$, where $\boldsymbol{f} = (f(x_1), ..., f(x_n))'$ and L is known as the graph Laplacian matrix, defined as $L = D - W$, where D is the diagonal matrix with $d_{ii} = \sum_{j=1}^{n} w_{ij}$.

Manifold regularization (Belkin *et al.* 2006; Seok, 2010) generalizes graph-based SSL with a regularized risk minimization framework. Let H be a Reproducing Kernel Hilbert Space (RKHS) of a kernel K. Manifold regularization obtains the discriminant function by solving

$$\min_{f \in H} \sum_{i=1}^{l} V(y_i, f(x_i)) + \lambda_1 ||f||_H^2 + \lambda_2 \boldsymbol{f}'L\boldsymbol{f} \tag{2.2}$$

where $V()$ is an arbitrary loss function, e.g., the hinge loss $|y_i - f(x_i)|_+$ for support vector machines (SVM), or squared loss for regularized least squares (RLS). The first two terms in (2.2) are the same as in supervised learning, while the third term is the additional regularization term for graph-based SSL.

3. Semi-supervised learinng using similarity and dissimilarity

From (2.1) and (2.2), we know that the existing methods cannot easily handle dissimilarity, which is the requirement that two items have different labels. A small weight or zero weight w_{ij} does not represent dissimilarity between x_i and x_j. In fact, a zero edge weight means no preference at all. A negative weight $w_{ij} < 0$ encourages a large difference between $f(x_i)$ and $f(x_j)$, but this creates several problems. First f needs to be bounded or $\{-\infty, \infty\}$ will be a trivial minimizer. Second, any negative weight in W will make (2.1) and ultimately the whole SSL problem non convex. It is highly desirable to keep the optimization problem convex. For these reasons, Goldberg *et al.* (2007) proposed a manifold regularization method with dissimilarity. The key idea of their study is to encode dissimilarity between x_i and x_j as $w_{ij}(f(x_i) + f(x_j))^2$. This term is zero if $f(x_i)$ and $f(x_j)$ have the same absolute value but opposite signs, thus encouraging different labels. The trivial case $f(x_i) = f(x_j) = 0$ is avoided by competing terms in the risk minimization framework (2.2). The weight w_{ij} remains positive, and represents the strength of the dissimilarity edge. They represented these ideas as a penalty term like (2.2) as

$$\frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f(x_i) - s_{ij}f(x_j))^2 \tag{3.1}$$

where $s_{ij} = 1$ if there is a similarity edge between x_i and x_j , $s_{ij=} -1$ if there is a dissimilarity edge between x_i and x_j . We can use (3.1) instead of (2.1) as the penalty term in (2.2). However, we should have prior domain knowledge s_{ij} to use it. In addition to s_{ij} being noisy, knowing s_{ij} for $i, j = 1, ..., n$ means that this data is no longer unlabeled. Furthermore, as explained in the Introduction, much of the data is unlabeled therefore, we develop an SSL using dissimilarity and similarity which can be used without prior information.

In this paper we intended the function to have the property such that $f(x_i) \approx f(x_j)$ if $w_{ij} \approx 1$ and $sign(f(x_i)) \neq sign(f(x_j))$ if $w_{ij} \approx 0$. The penalty term in our method is obtained as follows:

1. Calculate $W = (w_{ij})_{n\times n}$.
2. Let $v_{ij} = 2(w_{ij} - t)$, $0 < t < 1$. t : threshold
3. $E_{ij} = -v_{ij}(f(x_i) - f(x_j))^2$, if $v_{ij} <= 0$, $E_{ij} = v_{ij}(f(x_i) + f(x_j))^2$, elsewhere.
4. Penalty term: $P = \sum_{i,j=1}^{n} E_{ij}$

We can divide penalty term P into a similarity penalty term and a dissimilarity term as follows:

$$\begin{aligned} P &= \sum_{i,j=1}^{n} |v_{ij}|(f(x_i) - sign(v_{ij})f(x_j))^2 \\ &= \sum_{i,j=1}^{n} v_{ij}^{P}(f(x_i) - f(x_j))^2 + \sum_{i,j=1}^{n} v_{ij}^{N}(f(x_i) + f(x_j))^2 \\ &= 2\boldsymbol{f}'L^P\boldsymbol{f} + 2\boldsymbol{f}'L^N\boldsymbol{f} \\ &= \text{similarity penalty term} + \text{dissimilarity penalty term} \end{aligned} \tag{3.2}$$

where $L^P = D^P - V^P$, $L^N = V^N + D^N$, $V^P = (v_{ij}^P)_{n\times n}, V^N = (v_{ij}^N)_{n\times n}$, $v_{ij}^P = \max(v_{ij}, 0), v_{ij}^N = -\min(v_{ij}, 0)$ $D^P = diag(\sum_{i=1}^{n} v_{ij}^P)$ and $D_N = diag(\sum_{i=1}^{n} v_{ij}^N)$.

The objective function analog of (2.2) with (3.1) is given as

$$\min_{f\in H}\sum_{i=1}^{l} V(y_i, f(x_i)) + \lambda_1||f||_H^2 + 2\boldsymbol{f}'(\lambda_2 L^P + \lambda_3 L^N\ )\boldsymbol{f} \tag{3.3}$$

where λ_1 is a parameter which controls the complexity of the function in the ambient space while λ_2 and λ_3 control the complexity of the function in the intrinsic space (Belkin *et al.*, 2005). Also they control the strength of emphasis on similarity and dissimilarity respectively. If $\lambda_3 = 0$, the objective function (3.2) is the same as the manifold regularization objective function (2.2). Thus the proposed method is a more general method.

A modified version of the support vector machine (SVM) originally introduced by Vapnik (1995) in a least squares sense has been proposed for classification in Suykens (2000). The solution is given by a linear system instead of a quadratic programming problem. The fact that the LS-SVM has explicit primal-dual formulations has many advantages. Kernel tricks are used in the LS-SVM to treat the nonlinear relation between input variables and output variable. See Hwang (2010), Hwang (2008) and Shim *et al.* (2009) for details.

Here, we use the LS-SVM $\boldsymbol{f} = K\boldsymbol{\alpha}$ as the estimator in (3.3), where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)$, $K = (k_{ij})_{n \times n}$ an kernel matrix, and squared loss V. From (3.3) and $\boldsymbol{f} = K\boldsymbol{\alpha}$, objective function (3.3) can be represented as

$$\min_{\boldsymbol{\alpha}} (JK\boldsymbol{\alpha} - Y)'(JK\boldsymbol{\alpha} - Y) + \lambda_1 \boldsymbol{\alpha}' K \boldsymbol{\alpha} + \boldsymbol{\alpha}' K L^I K \boldsymbol{\alpha} \tag{3.4}$$

where $J = diag(1, ..., 1, 0, ..., 0)$ with the first l diagonal entries are 1 and the rest are 0, and Y is an $n \times 1$ dimensional label vector given by $Y = (y_1, ..., y_l, 0, ..., 0)'$ and $L^I = \lambda_2 L^P + \lambda_3 L^N$. The minimizer of (3.4) is

$$\boldsymbol{\alpha} = (JK + \lambda_1 I + L^I K)^{-1} Y \tag{3.5}$$

where I is an $n \times n$ identity matrix. From (3.5) we can estimate the labels as $sign(\boldsymbol{f}) = sign(K\boldsymbol{\alpha})$.

4. Hyperparameters selection

There are many things to be decided including hyperparameters which characterize the structure of the estimator of labels. They include the k value of $k - nn$ (k nearest neighborhood) or the value in the $\epsilon - ball$ method to construct the adjacency matrix, the graph distance function (Euclidean distance or cosine), the graph weight type (binary, distance or heat), the graph-weight parameter in the heat weight type, kernel type (radial basis function (rbf), linear, or polynomial), and kernel parameter. In this paper, we focus on the selection of regularization parameters λ_1, λ_2, λ_3 and the kernel parameter.

In the estimator, we should find the optimal values of regularization parameters λ_1, λ_2, and λ_3 and the kernel parameter σ^2. To select the parameters we define a cross validation (CV) function as

$$CV(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{f}_{\theta}^{(-i)}(x_i))^2, \tag{4.1}$$

where $\boldsymbol{\theta}$ is the set of hyperparameters and $\widehat{f}_{\boldsymbol{\theta}}^{(-i)}(x_i)$ is the estimated value of Y_i obtained from data without an i th observation. Since for each candidates of hyperparameters, $\widehat{f}_{\boldsymbol{\theta}}^{(-i)}(x_i)$ for $i = 1, \cdots, n$, should be evaluated, selecting parameters using the CV function is computationally formidable. By leaving-out-one lemma (Kimeldorf and Wahba, 1971) and the first order Taylor expansion, we have a generalized cross validation (GCV) function,

$$GCV(\boldsymbol{\theta}) = \frac{n \sum_{i=1}^{n} (Y_i - \widehat{f}_{\theta}^{(-i)}(x_i))^2}{(n - trace(\boldsymbol{S}))^2}. \tag{4.2}$$

where $\boldsymbol{S}$ is the hat matrix obtained from the linear equation (3.4) such that $\widehat{f} = SY$.

5. Numerical experiments

In this section, we empirically demonstrate the benefits of incorporating dissimilarity with two typical types of data - 2moons data and concentric circles data. The data sets have two categories and are shown in Figure 5.1. To reduce the complexity of model selection, some hyperparameters are fixed as in Sindhwani *et al.* (2005) and Goldberg *et al.* (2007). LapRLS (Belkin*et al.*, 2006) from (2.2) is used for comparison with the proposed method. We compute the average error rate (AER) and standard deviation of error rate (SDER) for 100 replications.

In all of the experiments, the RBF kernel $k_{ij} = \exp(-(x_i - x_j)^2/\sigma^2)$, $k - nn$, an adjacency matrix, and a Euclidean distance weight matrix were used. The graph-weight type and k values in the adjacency matrix were determined through a pilot experiment.

To investigate the benefits of incorporating dissimilarity we carried out our experiments with various sample size pairs $(l, u) = (2, 20), (4, 20), (6, 20), (8, 20)$, $(2, 200), (10, 200)$, (20, 200) for 2moons data and $(l, u) = (2, 200), (10, 200), (20, 200)$ for concentric circles data. An exact $l/2$ labeled sample and a $u/2$ unlabeled sample in each category were used in the experiments.

Through the pilot experiments, we know that $n-1$ nearest neighborhoods in the adjacency matrix and the distance-type weight matrix for our proposed method and the 6 nearest neighborhoods and heat kernel with 0.2 for LapRLS are suitable for 2moons data. For concentric circles data, nearest neighborhoods in the adjacency matrix and the binary-type weight matrix are suitable for the two methods. The calculated results are shown in Table 5.1. The table shows the AER and SDER in parenthesis.

Table 5.1 shows that LapRLS has lower AER and SDER when $l = 2, u = 20$ for 2moons data and when $l = 2, u = 200$ for concentric circles data. However, the proposed method shows better performance for other cases. This means that incorporating dissmilarity produces satisfactory results. The AER and SDER decrease as u increases, which is a desired property.

List of frequently selected parameter values are shown in Table 5.2. Since the size of the label data does not influence the values of the hyperparameters we enumerated the unlabeled data size only. From this table, we can conclude that similarity and dissimarity penalty terms are used properly, because one of the regularization parameters which controls similarity strength λ_2 tends to very small values for 2moons data and λ_3 tends to very small values for concentric circles data. Also we know that as the distribution changes, the parameters vary widely. Similar kernel parameter values were selected for the proposed method and LapRLS regardless of the distribution and sample size.

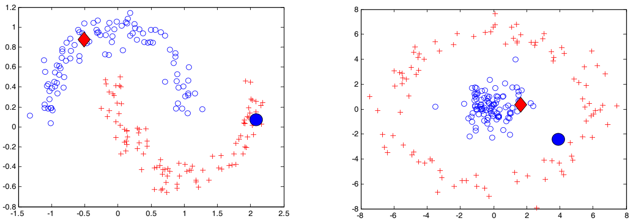

Figure 5.1 2moons data and concentric circles data of $l = 2$ and $u = 200$

Table 5.1 AER and SDER (parenthesis) of the proposed method and LapRLS. Hyperparameters chosen frequently from GCV are listed.

data set name	sample size (l, u)	Proposed method	LapRLS
2moons	(2, 20)	0.1581 (0.1262)	0.1211 (0.1189)
	(4, 20)	0.1123 (0.1179)	0.1129 (0.1180)
	(6, 20)	0.0504 (0.0747)	0.0637 (0.0799)
	(8, 20)	0.0410 (0.0662)	0.0671 (0.0886)
	(2, 200)	0.0676 (0.1084)	0.1091 (0.1261)
	(10, 200)	0.0106 (0.0292)	0.0369 (0.0496)
	(20, 200)	0.0079 (0.0227)	0.0296 (0.0448)
concentric circles	(2, 200)	0.0342 (0.0693)	0.0260 (0.0664)
	(10, 200)	0.0025 (0.0134)	0.0107 (0.0488)
	(20, 200)	0.0007 (0.0020)	0.0026 (0.0107)

Table 5.2 List of frequently selected parameter values

data set name	sample size (u)	parameter	Proposed method	LapRLS
2moons	20	σ^2	0.2	0.2
		λ_1	0.05	0.05
		λ_2	0.00001	10
		λ_3	0.8	
	200	σ^2	0.3	0.3
		λ_1	0.05	0.05
		λ_2	0.00005	50
		λ_3	0.8	
concentric circles	200	σ^2	0.5	0.5
		λ_1	0.1	0.005
		λ_2	50	5
		λ_3	0.0001	

6. Conclusions

In this paper, we dealt with a semi-supervised learning algorithm incorporating similarity and dissimilarity penalty terms. Through the experiments we showed that the proposed method derives satisfying results on semi-supervised classification problem. We knew that similarity and dissimarity penalty terms are used properly and the distribution have a big effect on parameter values. We also found that the propsed method has an advantage of using model selection methods such as GCV function.

References

Belkin, M., Sindhwani, V. and Niyogi, P. (2005). On manifold regularization. *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics.*

Belkin, M., Sindhwani, V. and Niyogi, P. (2006). Manifold regularization; A geometric framework for learning from examples, *Journal of Machine Learning Research*, **7**, 2329-2434.

Chapelle, O., Zien, A. and Scholkopf, B. (2006). *Semi-supervised learning*, MIT press.

Goldberg A., Zhu, X. and Wright, S. (2007). Dissmilarity in graph based semi-supervised classification. *Proceedings of the 10th International Conference of Artificial Intelligence and Statistics.*

Hwang, C. (2008). Mixed effect kernel binomial regression. *Journal of Korean Data and Information Science Society*, **19**, 1327-1334.

Hwang, H. T. (2010). Fixed size LS-SVM for multiclassification problems of large data sets. *Journal of Korean Data and Information Science Society*, **21**, 561-567.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and its Applications*, **33**, 82-95.

Lafferty, J. and Wasserman, L. (2007). Statistica analysis of semi-supervised regression, in *'NIPS'*.

Niyogi, P. (2008). *Manifold regularization and semi-supervised learning: Some throretical analysess*, Technical Report TR-2008-01, CS Dept, U. of Chicago.

Seok, K.H. (2010). Semi-supervised classification with LS-SVM formulation. *Journal of Korean Data and Information Science Society*, 461-470.

Shim, J., Park, H. and Hwang, C. (2009). A kernel machine for estimation of mean and volatility functions. *Journal of Korean Data and Information Science Society*, **20**, 905-912.

Sindhwani, V., Niyogi, P. and Belkin, M. (2005). Beyond the point cloud: from transductive to semisupervised learning. In *ICML05, 22nd International Conference on Machine Learning*.

Singh, A., Nowak, R. and Zhu, X. (2008). Unlabeled data: Now it helps, now it doesn't, in *'NIPS'*.

Suykens, J.A.K. (2000). Least squares support vector machine for classification and nonlinear modeling. *Neural Network World, Special Issue on PASE* 2000, **10**, 29-48.

Vapnik, V. (1995). *The nature of statistica learning theory*, Springer-Verlag, New York.

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

Zhu, X. and Goldberg, A. (2009). *Introduction to semi-supervised learning*, Morgan & Claypool.

Zhu, X., Ghahramani, Z. and Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic function, in *'ICML'*