AN INCREMENTAL ALGORITHM ABOUT THE AFFINITY-RULE BASED TRANSDUCTIVE LEARNING MACHINE FOR SEMI-SUPERVISED PROBLEM

Weijiang Long[1], Fengfeng Zhu[2] and Wenxiu Zhang[1]
[1)] *Institute of Information and Systems, Faculty of Sciences, Xi'an Jiaotong University, Xi'an 710049*
[2)] *Department of Appled Mathematics, South China University of Technology, Guangzhou 510640*

Abstract: One of the central problems in machine learning is how to effectively combine unlabelled and labelled data to infer the labels of unlabelled ones. In recent years, there has a growing interest on the transduction method. In this article, the transductive learning machines are described based on a so-called affinity rule which comes from the intuitive fact that if two objects are close in input space then their outputs should also be close, to obtain the solution of semi-supervised learning problem. By using the analytic solution for this problem, an incremental learning algorithm adapting to on-line data processing is derived.

Key words: Semi-supervised learning, Transductive learning machine, Support vector machines, Affinity measure, Incremental learning algorithm

1. INTRODUCTION

One of the most important subjects in current data mining research is semi-supervised learning in which some of the observations have been labelled by the supervisor, while the labels of others are not obtained for various reasons. We respectively call these two kinds of observations the labelled data and unlabelled data. The main problem to study is how to infer the proper label of the unlabelled data using the observations including labelled and unlabelled data and relevant knowledge. The classical method

for solving this problem is so-called induction-deduction method in which the labelled data are first analyzed to find a generalized rule and regard this rule to be justified for future observations (that is, from particularity to generality), and then this general rule is applied to the unlabelled data to infer their labels (that is, from generality to particularity).

In recent years, however, the transductive method, proposed by Vapnik(1998) has gained much concern. For semi-supervised learning, a general rule which is applied to both the unlabelled data and the possible other observations is indeed unnecessary in that only labeling those particular observations of unlabelled data is what we concern. The transduction method combined the labelled with the unlabelled data are used to derive the rule from particularity to particularity.

Up to now, there have been several examples of successful realization of transduction and experimentation on its superiority against traditional method. Chapell et al. (1999) implemented transductive inference by minimizing the leave-one-out error of ridge regression, and demonstrated that this transductive way for estimating values of the regression is more accurate than the traditional method. Bennettet et al.(1998) introduced semi-supervised support vector machines (S^3VM) by overall risk minimization, and demonstrated that S^3VM either make an improvement or show no significant difference in generalization compared to the usual structural risk minimization approach. Joachims (1999) suggested transductive support vector machines(TSVM) to deal with text classification. Furthermore, he presented a new transductive learning method in (Joachims, 2003) which can obtain the globally optimal solution by spectral methods. An algorithm was also proposed to robust achieve good generalization performance. Recently, Zhou et al.(2003a) studied semi-supervised learning problems by Hamilton method, and he latterly used objects programming involving norm in Zhou et al.(2003b) to derive the solution. However, it should be pointed out that is different from our paper. Zhou Guang-ya et al. (1993) gave many affinity-measures and applying rules which can be used in our discussing the machine learning problems.

In this paper, we deal with the semi-supervised learning transduction method based on affinity-rule by measuring the affinity of different objects in general space, and obtain an incremental learning algorithm. The principle comes from an intuitive fact that similar objects should have similar outputs. Because of the loose assumptions, this method has wider applications, more concise solution. We derive incremental learning algorithm adapting to on-line data processing. Its results are simple in expression and easy in calculation.

2. METHODS AND PROPERTIES

Let X^* be the input space and Y^* be the output space. The obtained data set is

$\{(\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2),\cdots,(\mathbf{x}_l,\mathbf{y}_l),\mathbf{x}_{l+1},\cdots,\mathbf{x}_{l+u}\}$,

$\mathbf{x}_i \in X^*, 1 \le i \le l+u;\quad \mathbf{y}_i \in Y^*, 1 \le i \le l, n = l+u$,

where $L^* = \{(\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2),\cdots,(\mathbf{x}_l,\mathbf{y}_l)\}$ is labelled data set and $U^* = \{\mathbf{x}_{l+1},\cdots,\mathbf{x}_{l+u}\}$ unlabelled data set. $\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_l\}$ are labelled objects and $\mathbf{y}_i$ is the label of $\mathbf{x}_i$. We aim at inferring the labels of unlabelled data. For convenience, we assume that the labels stand for the different classes and discuss the classification problem. Let the number of the classes be c and $\mathbf{e}_i$ be the vector with the i-th element being 1 and others being 0. Therefore Y^* may be taken as $\{\mathbf{e}_1,\mathbf{e}_2,\cdots,\mathbf{e}_c\}$ in classification problem. In order to obtain the labels estimates of $\mathbf{x}_i$ for $i>l$, a general labelling variable $\mathbf{z}_i \in R^c$ is often considered, where $\mathbf{z}_i = (z_{i1},z_{i2},\cdots,z_{ic})^T$ is an auxiliary variable to simplify the complexity in solving this problem, and $\mathbf{z}_i$ is not necessary to hold the same form as $\mathbf{e}_j$. Then we use $\mathbf{y}_i = \mathbf{e}_{\arg\max\{z_{ik};1\le k\le c\}}$ as the estimates of the labels. The following conditions are requested for $\mathbf{z}_i$. 1) If $\mathbf{x}_i$ is close to $\mathbf{x}_j$, then the general label $\mathbf{z}_i$ is also close to $\mathbf{z}_j$. 2) For the labelled data, the $\mathbf{z}_i$ and $\mathbf{y}_i$ are as close as possible. The measure of the closeness for two objects can be taken as various modes including the inclusion degree like that in Zhang et al.(1996), similarity (or dissimilarity, distance) etc.. The former measures have wider applications. For example, they can be used for the general symbolic data in the cases where even the symmetry does not hold. We use $s_{ij} = s(\mathbf{x}_i,\mathbf{x}_j)$ to measure the affinity between two input objects $\mathbf{x}_i$ and $\mathbf{x}_j$ where the matrix (s_{ij}) need not to be symmetric, but assume them to be positive. If the entries of (s_{ij}) are not all positive, we may consider $s_{ij}^* = s_{ij} + c$, $s > -\min\{s_{ij};1 \le i,j \le n\}$. Let $t_{ij} = t(\mathbf{z}_i,\mathbf{z}_j)$ be the affinity-measure between two outputs $\mathbf{z}_i$ and $\mathbf{z}_j$, and *f(v), g(v), h(v)* be nonnegative increasing functions of v. The idea based on the affinity-rule is that the greater the affinity-measure between $\mathbf{x}_i$ and $\mathbf{x}_j$, the smaller the $-f(s_{ij})$ as the dis-affinity-measure between $\mathbf{x}_i$ and $\mathbf{x}_j$, and the smaller the $g(t_{ij})$ dis-affinity-measure between $\mathbf{z}_i$ and $\mathbf{z}_j$; $t(\mathbf{z}_i,\mathbf{y}_i)$ should be as small as possible for $1 \le i \le l$. Therefore, we obtain a general framework as follows

$$\begin{aligned} \max_{\mathbf{z}_1,\mathbf{z}_2,\cdots,\mathbf{z}_n,\xi} \quad & -\sum_{i=1}^{n}\sum_{j=1}^{n} f(s(\mathbf{x}_i,\mathbf{x}_j))g(t(\mathbf{z}_i,\mathbf{z}_j)) - Ch(\xi), \\ s.t. \quad & \frac{1}{l}\sum_{i=1}^{l} r(t(\mathbf{z}_i,\mathbf{y}_i)) \le \xi, \quad \xi \ge 0, \end{aligned}$$

where C is a penalty factor which takes a tradeoff role among the multi-objects functions.

As a concrete realization, we take the decreasing function of the squared distance between two objects as the affinity-measure, and $\| \mathbf{p}-\mathbf{q} \|^2$ as the dis-affinity-measure. Let $f(v)=h(v)=g(v)=r(v)=v$, $t(\mathbf{p},\mathbf{q}) = \| \mathbf{p}-\mathbf{q} \|^2$, $\mathbf{z}_i = (z_{i1}, z_{i2}, \cdots, z_{ic})^T$, $Z = (\mathbf{z}_1^T, \mathbf{z}_2^T, \cdots, \mathbf{z}_n^T)^T = (\mathbf{z}_{(1)}, \mathbf{z}_{(2)}, \cdots, \mathbf{z}_{(c)})$, $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{ic})^T$, $1 \le i \le l$; $\mathbf{y}_j = \mathbf{0}$, $j \ge l+1$, and $Y_0 = (\mathbf{y}_1^T, \mathbf{y}_2^T, \cdots, \mathbf{y}_n^T)^T = (\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \cdots, \mathbf{y}_{(c)})$. We always suppose $s_{ij} > 0$ without special declaration. Then we obtain the formal expression for semi-supervised transductive learning machine based on the affinity-rule as follows

$$\left.\begin{aligned} \min_{\mathbf{z}_1,\mathbf{z}_2,\cdots,\mathbf{z}_n,\xi} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij} \| \mathbf{z}_i - \mathbf{z}_j \|^2 + C\xi, \\ s.t. \quad & \frac{1}{l}\sum_{i=1}^{l} \| \mathbf{z}_i - \mathbf{y}_i \|^2 \le \xi, \quad \xi \ge 0, \end{aligned}\right\}. \qquad (P)$$

This is a convex programming and there is a globally optimal solution. Its Lagrangian function is

$$L = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij} \| \mathbf{z}_i - \mathbf{z}_j \|^2 + C\xi + \lambda\left(\frac{1}{l}\sum_{i=1}^{l} \| \mathbf{z}_i - \mathbf{y}_i \|^2 - \xi\right) - \mu\xi, \quad \xi \ge 0, \mu \ge 0.$$

Note that

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij} \| \mathbf{z}_i - \mathbf{z}_j \|^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij} \sum_{k=1}^{c} (z_{ik} - z_{jk})^2 \\ &= \sum_{k=1}^{c}\left[\sum_{i=1}^{n}\sum_{j\ne i} (s_{ij} + s_{ji}) z_{ik}^{\ 2} - 2\sum_{i=1}^{n}\sum_{j<i} (s_{ij} + s_{ji}) z_{ik} z_{jk}\right]. \end{aligned}$$

Let

$$w_{ij} = \frac{1}{2}(s_{ij} + s_{ji}),\ j \neq i; \quad w_{ii} = 0; \quad W = (w_{ij})_{n\times n};$$
$$d_i = \sum_{j\neq i} w_{ij},\ j \neq i;\ D = \mathrm{diag}(d_1, d_2, \cdots, d_n), \quad A = D - W;$$
$$I_l = \mathrm{diag}(1,1,\cdots,1,0,\cdots,0),$$

where there are l 1's in I_l, and A is a symmetric matrix. Therefore,

$$L_P = 2\sum_{k=1}^{c} \mathbf{z}_{(k)}^T A \mathbf{z}_{(k)} + \frac{\lambda}{l}\sum_{i=1}^{l} \| \mathbf{z}_i - \mathbf{y}_i \|^2 + (C - \lambda - \mu)\xi, \xi \geq 0, \mu \geq 0. \quad (1)$$

To obtain the *K-T* points of (*P*), we calculate the derivative

$$\left.\begin{aligned} \frac{\partial L}{\partial \xi} &= C - \lambda - \mu \\ \frac{\partial L}{\partial \mathbf{z}_{(k)}} &= 4A\mathbf{z}_{(k)} + \frac{2\lambda}{l} I_l(\mathbf{z}_{(k)} - \mathbf{y}_{(k)}) \end{aligned}\right\}.$$

When Z is the *K-T* points of (*P*), then

$$\left.\begin{aligned} C - \lambda - \mu = 0, \lambda \geq 0, \mu \geq 0, \\ (2lA + \lambda I_l)\mathbf{z}_{(k)} = \lambda I_l \mathbf{y}_{(k)}. \end{aligned}\right\} \quad (2)$$

and the *KKT* conditions are hold

$$\lambda(\frac{1}{l}\sum_{i=1}^{l} \| \mathbf{z}_i - \mathbf{y}_i \|^2 - \xi) = 0, \quad \mu\xi = 0. \quad (3)$$

In consideration of the limited space, we will omitt the following some lemmas and theorems, and prove them in another article.

Lemma 1. The solutions of (*P*) with the condition (2) are constant values when $\lambda = 0$, which are trivial solutions. If the number of the labels for the labelled data is greater than one, then $\lambda > 0$.

Lemma 2. When $\lambda > 0$, we have
1) $(2lA + \lambda I_l)$ is invertible, and
2) $\rho(2l(2lD + \lambda I_l)^{-1} W) < 1$.

We deal with the dual problem of (*P*) now. Substituting the results of (2) into (1), and using (2), we obtain $0 \le \lambda \le C$ and

$$L_D = -\frac{1}{l}\sum_{k=1}^{c} \lambda \mathbf{y}_{(k)}^T I_l \mathbf{z}_{(k)} + \frac{1}{l}\sum_{k=1}^{c} \lambda \mathbf{y}_{(k)}^T I_l \mathbf{y}_{(k)} .$$

So, when $0 < \lambda \le C$, by Lemma 2, the dual programming is

$$\left.\begin{array}{l} \max\limits_{\lambda} -\frac{1}{l}\sum\limits_{k=1}^{c} \lambda \mathbf{y}_{(k)}^T I_l (2lA + \lambda I_l)^{-1} \lambda I_l \mathbf{y}_{(k)} + \frac{1}{l}\sum\limits_{k=1}^{c} \lambda \mathbf{y}_{(k)}^T I_l \mathbf{y}_{(k)} \\ 0 < \lambda \le C \end{array}\right\}. \qquad (D).$$

Lemma 3. For $0 < \lambda \le C$, L_D reaches the maximum when $\lambda = C$.

Theorem 4. For the nontrivial solutions $(Z, \xi, \lambda, \mu) = (\hat{Z}, \hat{\xi}, C, 0)$ are the Lagrangian saddle-point of (*P*), and $(\hat{Z}, \hat{\xi})$ are the globally optimal solution of (*P*), where

$$\hat{Z} = C(2lA + CI_l)^{-1} Y_0 \quad \text{and} \quad \hat{\xi} = \frac{1}{l}\sum_{i=1}^{l} \| \hat{\mathbf{z}}_i - \mathbf{y}_i \|^2 .$$

3. MAIN RESULTS

To discuss the problem with real-time constrain, we study the situations with information increment. Suppose that the given $n=l+u$ data have been expressed with the notations in Section 2. Let $s = 2l/C$ and denote respectively by $A_n, I_l(n), Y_n$, and U_n as the aforementioned A, I_l, Y_0 and U in the case of n given data where $U_n = sA_n + I_l(n)$, $Z(n) = U_n^{-1} Y_n = (sA_n + I_l(n))^{-1} Y_n$, $I_l(n) = \text{diag}(1,1,\cdots 1,0,\cdots,0)$, and rearrange these n data in such a way that the labelled data are in the front of the unlabelled data. We add a star to $Z(n+1)$ to stand for the rearranging version of $Z(n+1)$ with $n+1$ data.

Now, we study the semi-supervised learning problem with information increment for different cases respectively. In consideration of the limited space, we will only give an outline of some proofs as follows.

3.1) The $(n+1)$-th point is unlabelled. Let

$$Y_{n+1} = \begin{pmatrix} Y_n \\ \mathbf{0} \end{pmatrix}, \quad Z(n+1) = \begin{pmatrix} Z_n \\ \mathbf{z}_{n+1}^T \end{pmatrix}, \quad I_l(n+1) = \begin{pmatrix} I_l(n) & 0 \\ 0 & 0 \end{pmatrix}_{(n+1)\times(n+1)},$$

and

$$A_{n+1} = \begin{pmatrix} A_n & A_{n,n+1} \\ A_{n+1,n} & d_{n+1} \end{pmatrix}.$$

where $A_{n,n+1}$ are an affinity vector between the $(n+1)$-th point and the n data given before, that is, $A_{n,n+1} = (a_{1,n+1}, a_{2,n+1}, \cdots, a_{n,n+1})^T$, $A_{n+1,n} = A_{n,n+1}^T$, and $d_{n+1} = \sum_{j \neq n+1} a_{n+1,j}$.

By the analytic solution of transductive learning machine based on the affinity-rule, we get

$$Z(n+1) = [sA_{n+1} + I_l(n+1)]^{-1} Y_{n+1}.$$

Then

$$sA_{n+1}Z(n+1) + I_l(n+1)[Z(n+1) - Y_{n+1}] = 0.$$

Therefore we have

$$\left.\begin{aligned} &sA_nZ_n + sA_{n,n+1}\mathbf{z}_{n+1}^T + I_l(n)(Z_n - Y_n) = 0 \\ &sA_{n,n+1}^T Z_n + sd_{n+1}\mathbf{z}_{n+1}^T = 0 \end{aligned}\right\}.$$

Thus

$$\left.\begin{aligned} Z_n &= [sA_n + I_l(n) - \frac{s}{d_{n+1}} A_{n,n+1}A_{n,n+1}^T Z_n]^{-1} Y_n \\ \mathbf{z}_{n+1}^T &= \frac{-A_{n,n+1}^T Z_n}{d_{n+1}} \end{aligned}\right\}.$$

By Sherman-Morrison-Woodbory formula
$(M + BCD^T)^{-1} = M^{-1} - M^{-1}B(C^{-1} + D^T M^{-1}B)^{-1} D^T M^{-1}$,
we can write some simple expressions in Theorem 5.

3.2) The $(n+1)$-th point is labelled, and its label is $\mathbf{y}_{n+1}$. Define an elementary matrix as

$$P(i,j) = I - (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T$$

Then $P(i,j) = P^{-1}(i,j)$. Let

$$P_{n+1} = P(n+1,n)P(n+1,n-1)\cdots P(n+1,l+2)P(n+1,l+1),$$

and $Y_{n+1} = (Y_n^T, y_{n+1})^T$. With the similar notations in subsection 3.1 and the method in Section 2, we have

$$P_{n+1}Z(n+1) = [sP_{n+1}A_{n+1}P_{n+1}^{-1} + I_l(n+1) + I_{(l+1)}]^{-1} P_{n+1}\begin{pmatrix} Y_n \\ \mathbf{y}_{n+1}^Y \end{pmatrix},$$

where $I_{(i+1)} = \mathbf{e}_{i+1}\mathbf{e}_{i+1}^T$. Therefore we have

$$Z(n+1) = [sA_{n+1} + I_l(n+1) + I_{(n+1)}]^{-1}\begin{pmatrix} Y_n \\ \mathbf{y}_{n+1}^Y \end{pmatrix}.$$

So,

$$\left.\begin{array}{l} sA_{n+1}Z_n + sA_{n,n+1}\mathbf{z}_{n+1}^T + I_l(n)(Z_n - Y_n) = 0 \\ sA_{n,n+1}^T Z_n + sd_{n+1}\mathbf{z}_{n+1}^T + \mathbf{z}_{n+1}^T - \mathbf{y}_{n+1}^T = 0 \end{array}\right\},$$

we obtain

$$\left.\begin{array}{l} Z_n = \left[sA_n + I_l(n) - \dfrac{s^2}{1+sd_{n+1}} A_{n,n+1}A_{n,n+1}^T\right]^{-1} \left[I_l(n)Y_n - \dfrac{s}{1+sd_{n+1}} A_{n,n+1}\mathbf{y}_{n+1}^T\right] \\ \mathbf{z}_{n+1}^T = -\dfrac{1}{1+sd_{n+1}}(\mathbf{y}_{n+1}^T - sA_{n,n+1}^T Z_n) \end{array}\right\}$$

Since

$$\begin{aligned} &[U_n - \frac{s^2}{1+sd_{n+1}} A_{n,n+1}A_{n,n+1}^T]^{-1} \\ &= U_n^{-1} - \frac{s^2}{1+sd_{n+1}}\left(1 - \frac{s^2}{1+sd_{n+1}} A_{n,n+1}^T U_n^{-1} A_{n,n+1}\right)^{-1} U_n^{-1} A_{n,n+1}A_{n,n+1}^T U_n^{-1} \\ &= \left[I - \frac{s^2}{1+sd_{n+1} - s^2 A_{n,n+1}^T U_n^{-1} A_{n,n+1}} U_n^{-1} A_{n,n+1}A_{n,n+1}^T\right] U_n^{-1}, \end{aligned}$$

we may give some simple formulae in Theorem 5.

According to the analytic expression for transductive learning machine based on the affinity-rule, we can adjust the result in such a way that the labelled data are arranged in the front of the unlabelled data. Then

$$Z^*(n+1) = P_{n+1}Z(n+1).$$

We now sumarize the cases of 3.1)- 3.2) in the following theorem.

Theorem 5. Suppose that the solution of transductive learning machine based on the affinity-rule for the n data has been given as $Z(n) = U_n^{-1} Y_n = (sA_n + I_l(n))^{-1} Y_n$. Let P_{n+1} be the products of some elementary matrices defined as above. The incremental learning solution is that

1) if the incremental information is the $(n+1)$-th data which is an unlabelled point,then

$$\left.\begin{aligned} Z_n &= [I + \frac{1}{d_{n+1} - sA_{n,n+1}^T U_n^{-1} A_{n,n+1}} U_n^{-1} A_{n,n+1} A_{n,n+1}^T] Z(n) \\ \mathbf{z}_{n+1}^T &= \frac{-A_{n,n+1}^T Z_n}{d_{n+1}} \end{aligned}\right\}.$$

and

$Z^*(n+1) = Z(n+1) = (Z_n^T, \mathbf{z}_{n+1})^T$

2)if the incremental information is the $(n+1)$-th data which is a labelled point and its label is $\mathbf{y}_{n+1}$,

$$\left.\begin{aligned} Z_n &= \left[I - \frac{s^2}{1 + sd_{n+1} - s^2 A_{n,n+1}^T U_n^{-1} A_{n,n+1}} U_n^{-1} A_{n,n+1} A_{n,n+1}^T \right] [Z(n) - \\ &\qquad \frac{s}{1 + sd_{n+1}} U_n^{-1} A_{n,n+1} \mathbf{y}_{n+1}^T] \\ \mathbf{z}_{n+1}^T &= -\frac{1}{1 + sd_{n+1}} (\mathbf{y}_{n+1}^T - sA_{n,n+1}^T Z_n) \end{aligned}\right\}.$$

When arranging the labelled data in the former part and the n abeled data in the latter part of the solution vector, we obtain

$Z^*(n+1) = P_{n+1} Z(n+1) = P_{n+1} (Z_n^T, \mathbf{z}_{n+1})^T$.

4. CONCLUSION

We have described a semi-supervised learning problems by the transduction method in which both the labelled and the unlabelled data are used to derive the rule from particularity to particularity. Our transductive learning machine based on affinity has several advantages. It is adaptable to

general input space objects and even can be extended to the situation with default values like those in Zhou et al. (1993). This method only requires a measure of affinity between the input objects and the input observation labels given by the supervisor to infer the proper labels on the other unlabelled objects. We have derived an incremental learning algorithm. Unlike other methods, when a new data comes, our method does not need to use both the original and the new data to solve the entire problem but uses the previous result and the new data to recurrence solution, and therefore is more adaptive to on-line processes.

REFERENCES

1. Ben-Hur, A. , Horn, D. , Siegelmann, H. T. & Vapnik V. (2001). Support vector clustering. *Journal of Machine Learning Research, 2*:125-137.
2. Bennett, K. , & Demiriz A. (1999). Semi-supervised support vector machines . In M. S. Kearns, S. A. Solla, & D. A. Cohn, (Eds.), *Advances in Neural Information Processing Systems 11* (pp.368-374), Cambridge: MIT Press.
3. Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2):121-167 .
4. Chapelle, O. , Vapnik, V. , & Weston J. (1999). Transductive inference for estimating values of functions. In Sara A. Solla, Todd K. Leen & Klaus-Robert M¨ller (Eds.), *Advances in Neural Information Processing Systems 12* (pp.421-427). Cambridge :MIT Press.
5. Chapelle, O. , Weston, J. , & Scholkopf, B. (2003). Cluster kernels for semi-supervised learning. In T. G. Diettrich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14*. Cambridge: MIT Press(in press).
6. Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *International Conferenceon Machine Learning(ICML)* (pp.200-209).
7. Joachims, T. (2003). Transductive learning via spectral graph partitioning, *Proceedings of the International Conference on Machine Learning (ICML).*
8. Vapnik, V.(1998). *Statistical learning theory*. New York: Wiley.
9. Zhou, D. , Bousquet, O. , Lal, T. N. , Weston, J. , & Scholkopf B. (2003a). Learning with local and global consistency. (112), *Max Planck Institute for Biological Cybernetics, Tuebingen, Germany* (June 2003).
10. Zhou, D. , Bousquet, O. , Lal, T. N. , Weston, J. , & Scholkopf B. (2003b). Learning with local and global consistency. In S. Thrun, L. Saul & B. Scholkopf (Eds.), *Advances in Neural Information Processing Systems 16.*, Cambridge: MIT Press (in press).
11. Zhang, W.X., & Leung, Y. (1996). *The principle of uncertainty inference*. Xi'an: Xi'an Jiaotong University Press. (in Chinese).
12. Zhou, G. Y., & Xia, L. X. (1993). *Non-metric data analysis and its applications*. Beijing: Science Press. (in Chinese).