Learning from Labeled and Unlabeled Data: Semi-supervised Learning and Ranking

Dengyong Zhou

zhou@tuebingen.mpg.de

Dept. Schölkopf, Max Planck Institute for Biological Cybernetics, Germany

Learning from Examples

- Input space $\mathcal{X}$, and output space $\mathcal{Y} = \{1, -1\}$.
- Training set $S = \{z_1 = (x_1, y_1), \ldots, z_l = (x_l, y_l)\}$ in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ drawn i.i.d. from some unknown distribution.
- Classifier $f : \mathcal{X} \to \mathcal{Y}$.

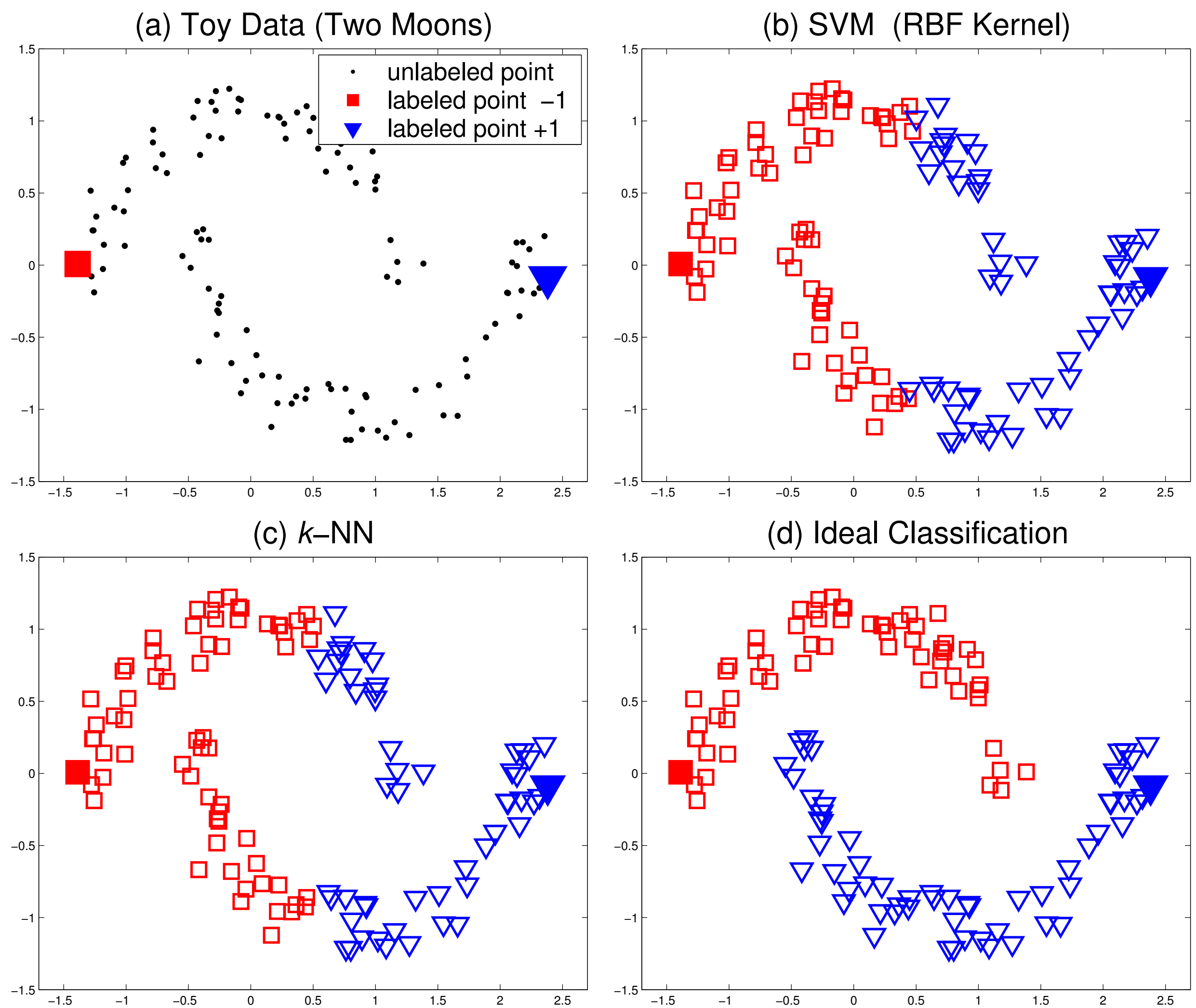
Transductive Setting

- Input space $\mathcal{X} = \{x_1, \ldots, x_n\}$, and output space $\mathcal{Y} = \{1, -1\}$.
- Training set $S = \{z_1 = (x_1, y_1), \ldots, z_l = (x_l, y_l)\}$.
- Classifier $f : \mathcal{X} \to \mathcal{Y}$.

Intuition about classification: Manifold

- **Local consistency**. Nearby points are likely to have the same label.
- **Global consistency**. Points on the same structure (typically referred to as a cluster or manifold) are likely to have the same label.

A Toy Dataset (Two Moons)



Algorithm

1. Form the affinity matrix W defined by $W_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$ if $i \neq j$ and $W_{ii} = 0$.
2. Construct the matrix $S = D^{-1/2} W D^{-1/2}$ in which D is a diagonal matrix with its (i, i)-element equal to the sum of the i-th row of W.
3. Iterate $f(t+1) = \alpha S f(t) + (1-\alpha) y$ until convergence, where α is a parameter in $(0, 1)$.
4. Let f^* denote the limit of the sequence $\{f(t)\}$. Label each point x_i as $y_i = \text{sgn}(f_i)$.

Convergence

Theorem. *The sequence* $\{f(t)\}$ *converges to* $f^* = \beta(I - \alpha S)^{-1} y$, *where* $\beta = 1 - \alpha$.

Proof. Suppose $F(0) = Y$. By the iteration equation, we have

$$f(t) = (\alpha S)^{t-1} Y + (1 - \alpha) \sum_{i=0}^{t-1} (\alpha S)^i Y. \quad (1)$$

Since $0 < \alpha < 1$ and the eigenvalues of S in $[-1, 1]$,

$$\lim_{t \to \infty} (\alpha S)^{t-1} = 0, \text{ and } \lim_{t \to \infty} \sum_{i=0}^{t-1} (\alpha S)^i = (I - \alpha S)^{-1}. \quad (2)$$

Regularization Framework

Cost function

$$\mathcal{Q}(f) = \frac{1}{2}\left[\sum_{i,j=1}^{n} W_{ij}\left(\frac{1}{\sqrt{D_{ii}}}f_i - \frac{1}{\sqrt{D_{jj}}}f_j\right)^2 + \mu\sum_{i=1}^{n}\left(f_i - y_i\right)^2\right]$$

- **Smoothness term.** Measure the changes between nearby points.
- **Fitting term.** Measure the changes from the initial label assignments.

Regularization Framework

Theorem. $f^* = \arg\min_{f \in \mathcal{F}} \mathcal{Q}(f)$.

Proof. Differentiating $\mathcal{Q}(f)$ with respect to f, we have

$$\left.\frac{\partial \mathcal{Q}}{\partial f}\right|_{f=f^*} = f^* - Sf^* + \mu(f^* - y) = 0, \tag{1}$$

which can be transformed into

$$f^* - \frac{1}{1+\mu} Sf^* - \frac{\mu}{1+\mu} y = 0. \tag{2}$$

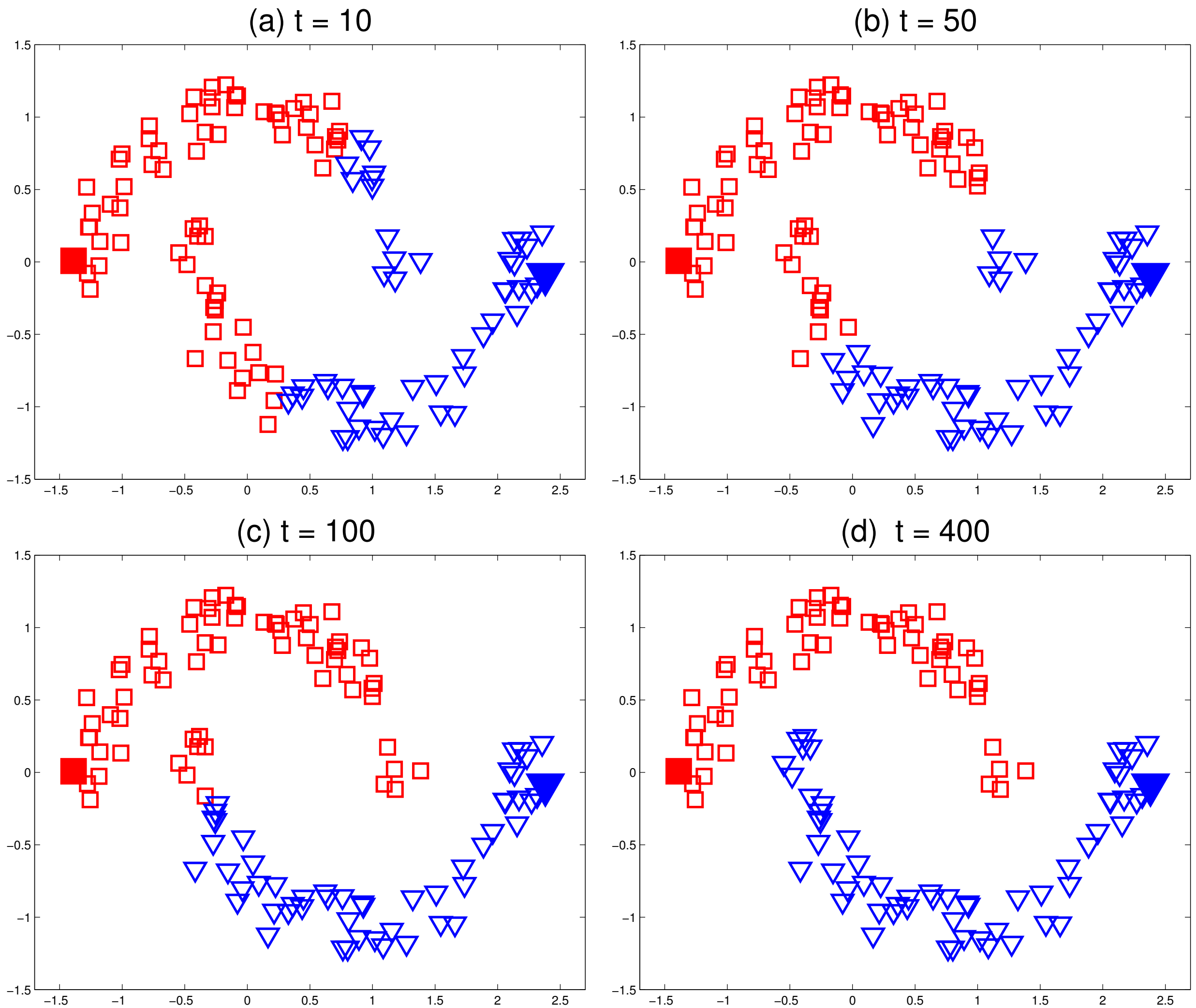
Let $\alpha = 1/(1+\mu)$ and $\beta = \mu/(1+\mu)$. Then

$$(I - \alpha S) f^* = \beta y. \tag{3}$$

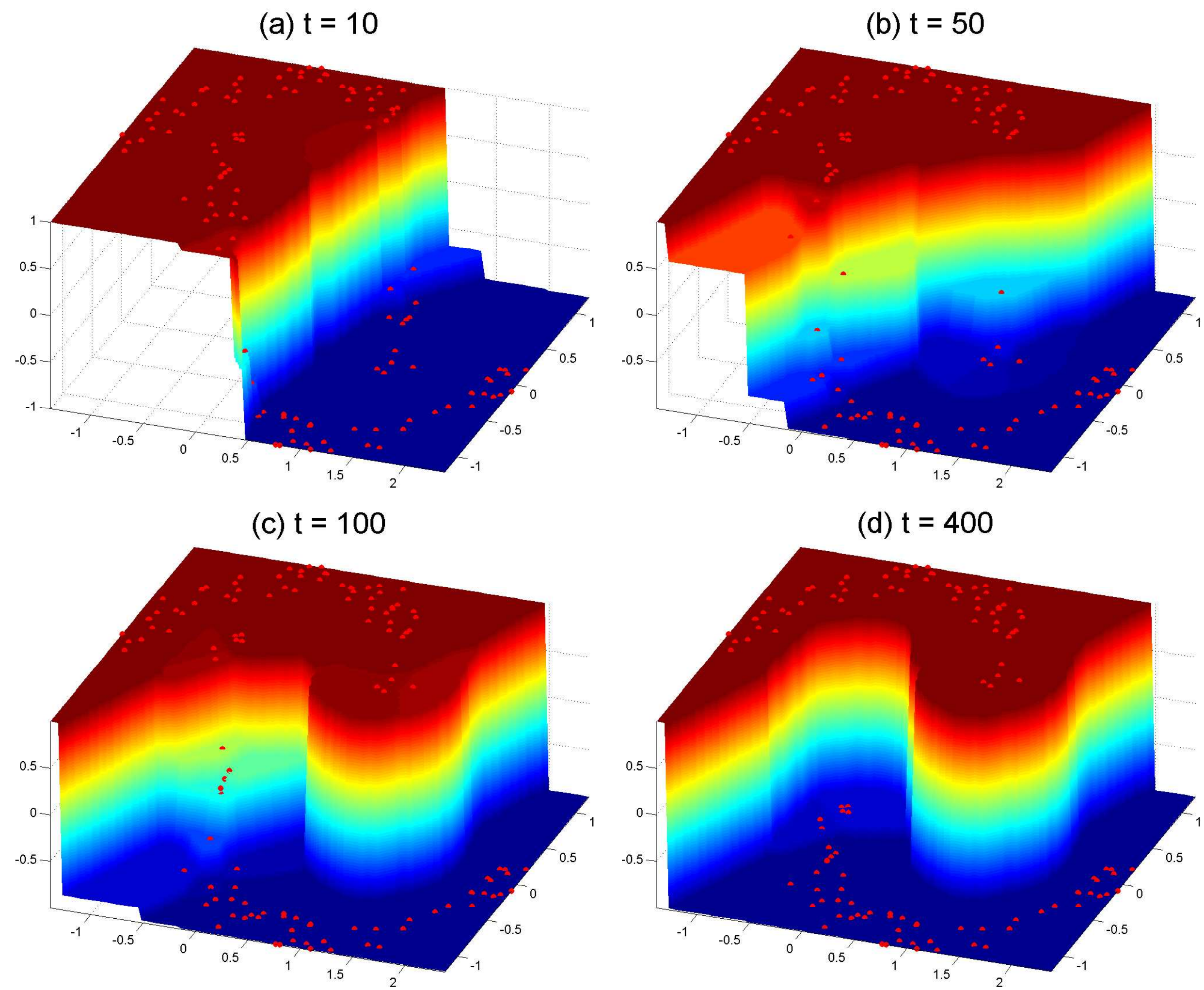
Two Variants

- Substitute $P = D^{-1}W$ for S in the iteration equation. Then $f^* = (I - \alpha P)^{-1}y$.
- Replace S with P^T, the transpose of P. Then $f^* = (I - \alpha P^T)^{-1}y$, which is equivalent to $f^* = (D - \alpha W)^{-1}y$.

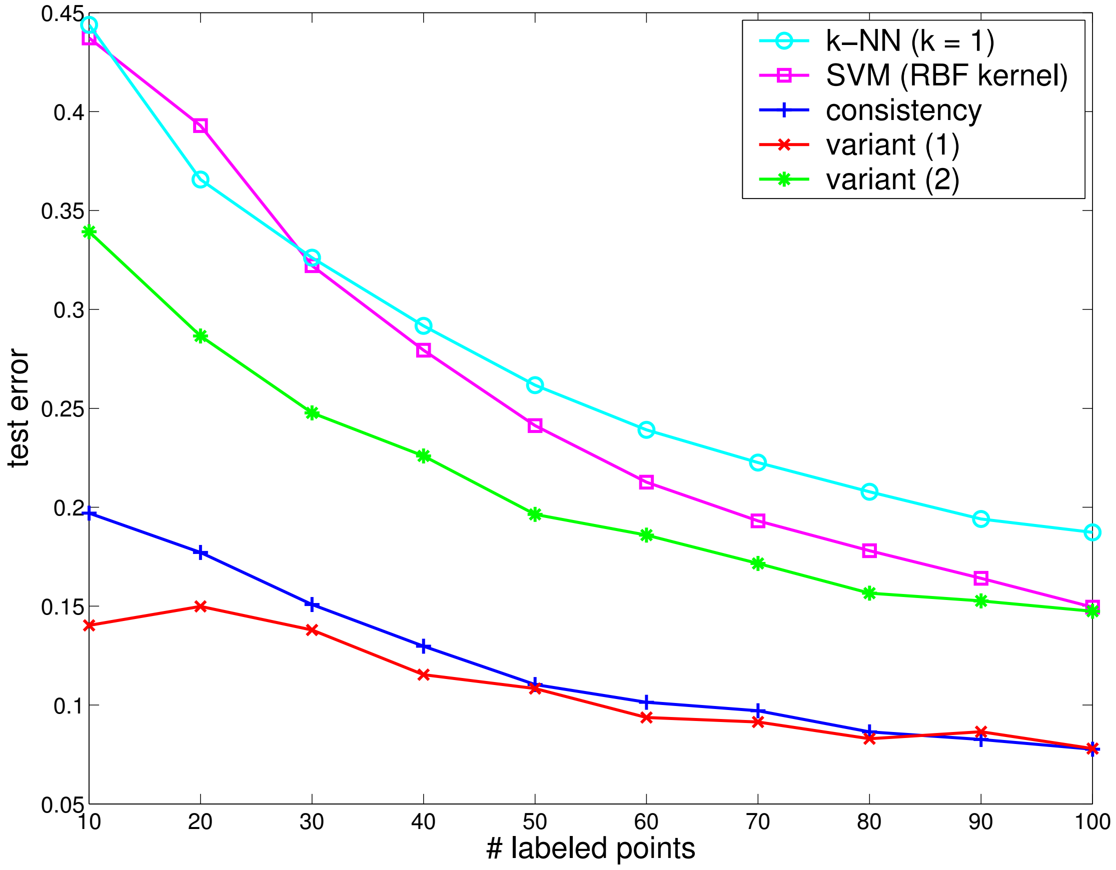
Toy Problem



Toy Problem

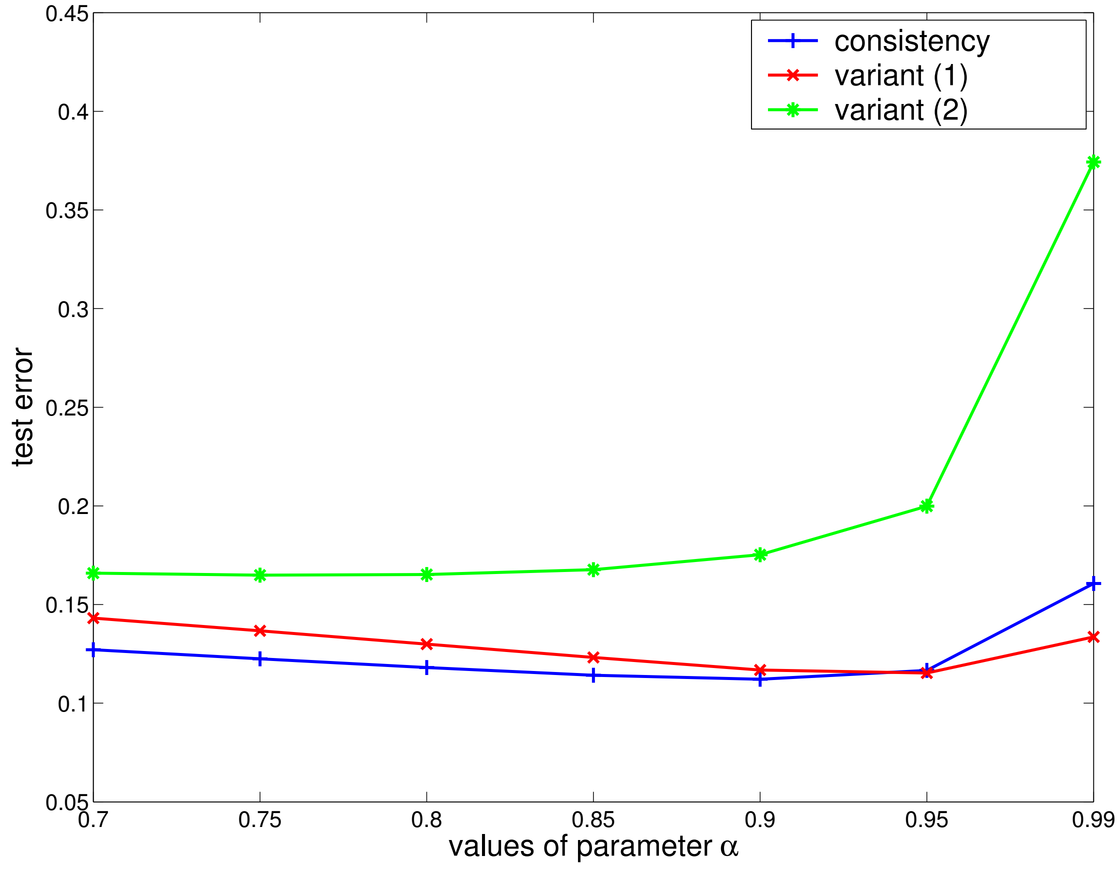


Handwritten Digit Recognition (USPS)



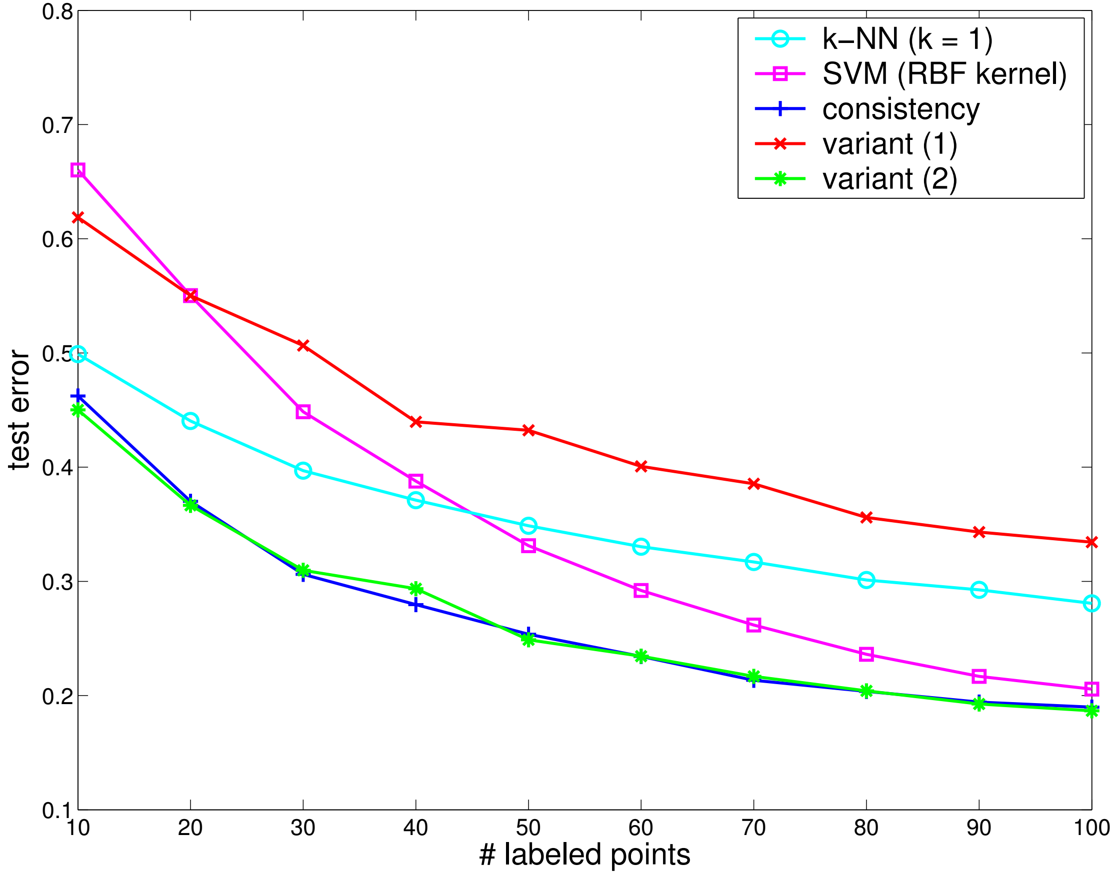
Dimension: 16x16. Size: 9298. ($\alpha = 0.95$)

Handwritten Digit Recognition (USPS)



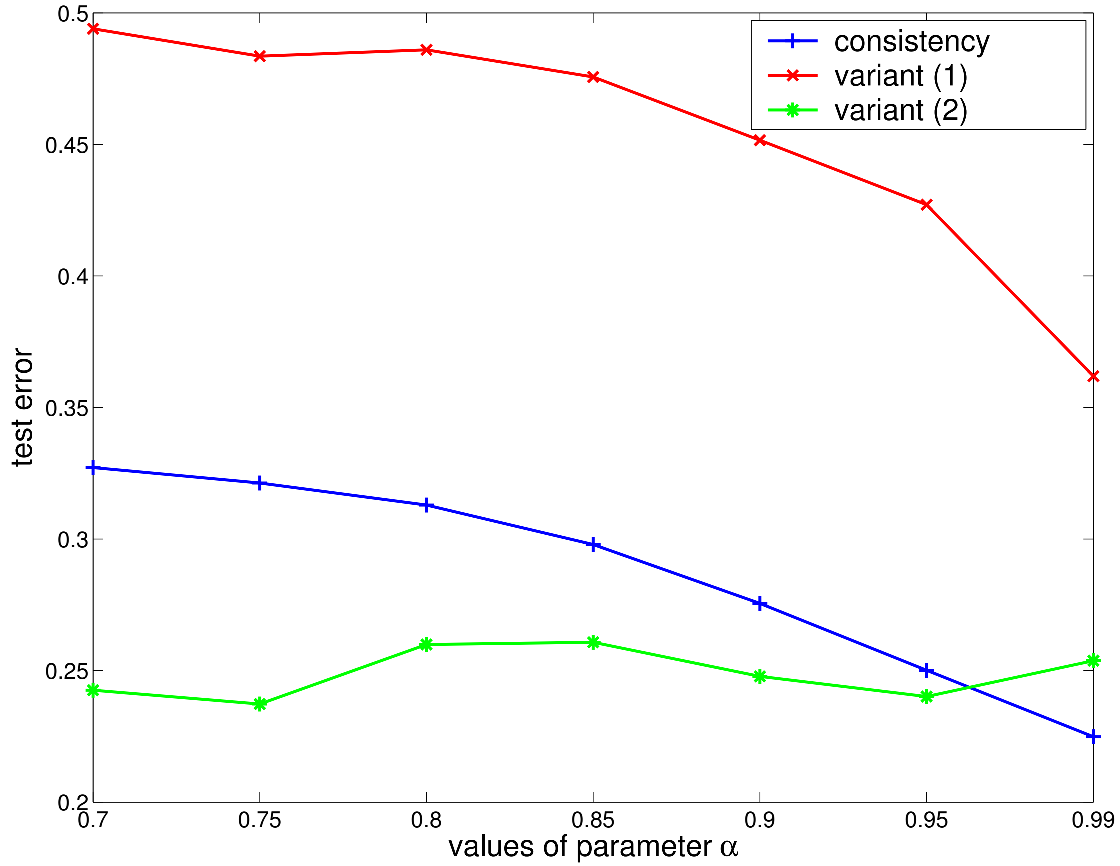
Size of labeled data: $l = 50$.

Text Classification (20-newsgroups)



Dimension: 8014. Size: 3970. ($\alpha = 0.95$)

Text Classification (20-newsgroups)



Size of labeled data: $l = 50$.

Spectral Graph Theory

Normalized graph Laplacian $\Delta = D^{-1/2}(D - W)D^{-1/2}$.
Linear operator on the space of functions defined on the Graph.

Theorem. $\sum_{i,j} W_{ij} \left(\frac{1}{\sqrt{D_{ii}}} f_i - \frac{1}{\sqrt{D_{jj}}} f_j \right)^2 = \langle f, \Delta f \rangle$.
Discrete analogy of Laplace-Beltrami operator on Riemannian Manifold which satisfies

$$\int_{\mathcal{M}} \|\nabla f\|^2 = \int_{\mathcal{M}} \Delta(f) f.$$

Discrete Laplace equation $\Delta f = y$.
Green's function $G = \Delta^{\dagger}$.

Reversible Markov Chains

Lazy random walk defined by the transition probability matrix $P^* = (1-\alpha)I + \alpha D^{-1}W, \alpha \in (0,1)$.

Hitting time $H_{ij} = E\{$ number of steps required for a random walk to reach a position x_j with an initial position $x_i\}$.

Commute time $C_{ij} = H_{ij} + H_{ji}$.

Theorem. *Let* $L = (D - \alpha W)^{-1}$. *Then*

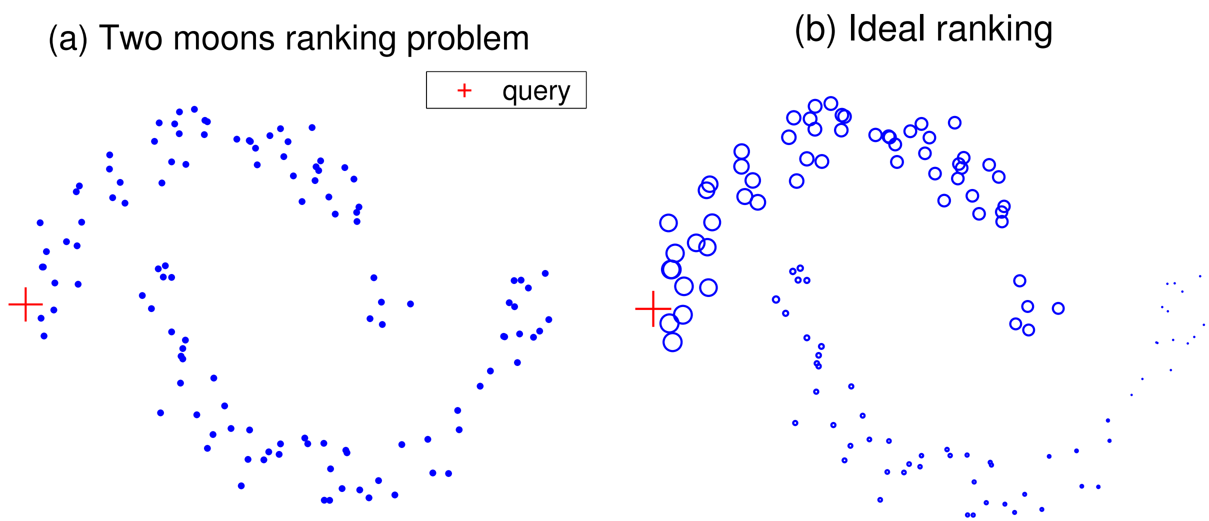
$$C_{ij} \propto L_{ii} + L_{jj} - L_{ij} - L_{ji}$$

Ranking Problem

Problem setting. Given a set of point $\mathcal{X} = \{x_1, ..., x_q, x_{q+1}, ..., x_n\} \subset \mathbb{R}^m$, the first q points are the queries. The task is to rank the remaining points according to their relevances to the queries.

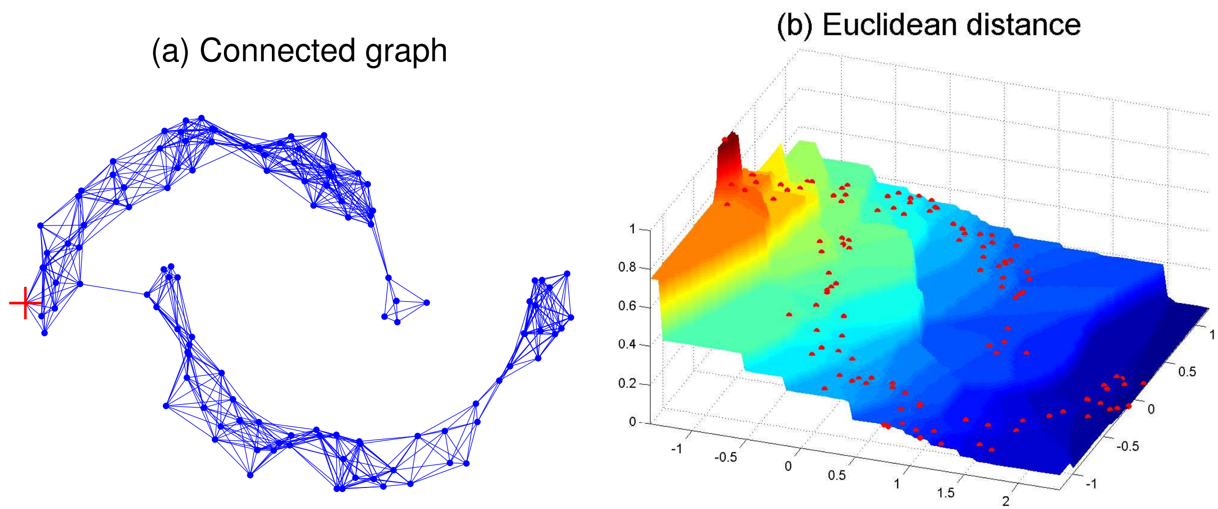
Examples. Image, document, movie, book, protein ("killer application"), . . .

Intuition of Ranking: Manifold



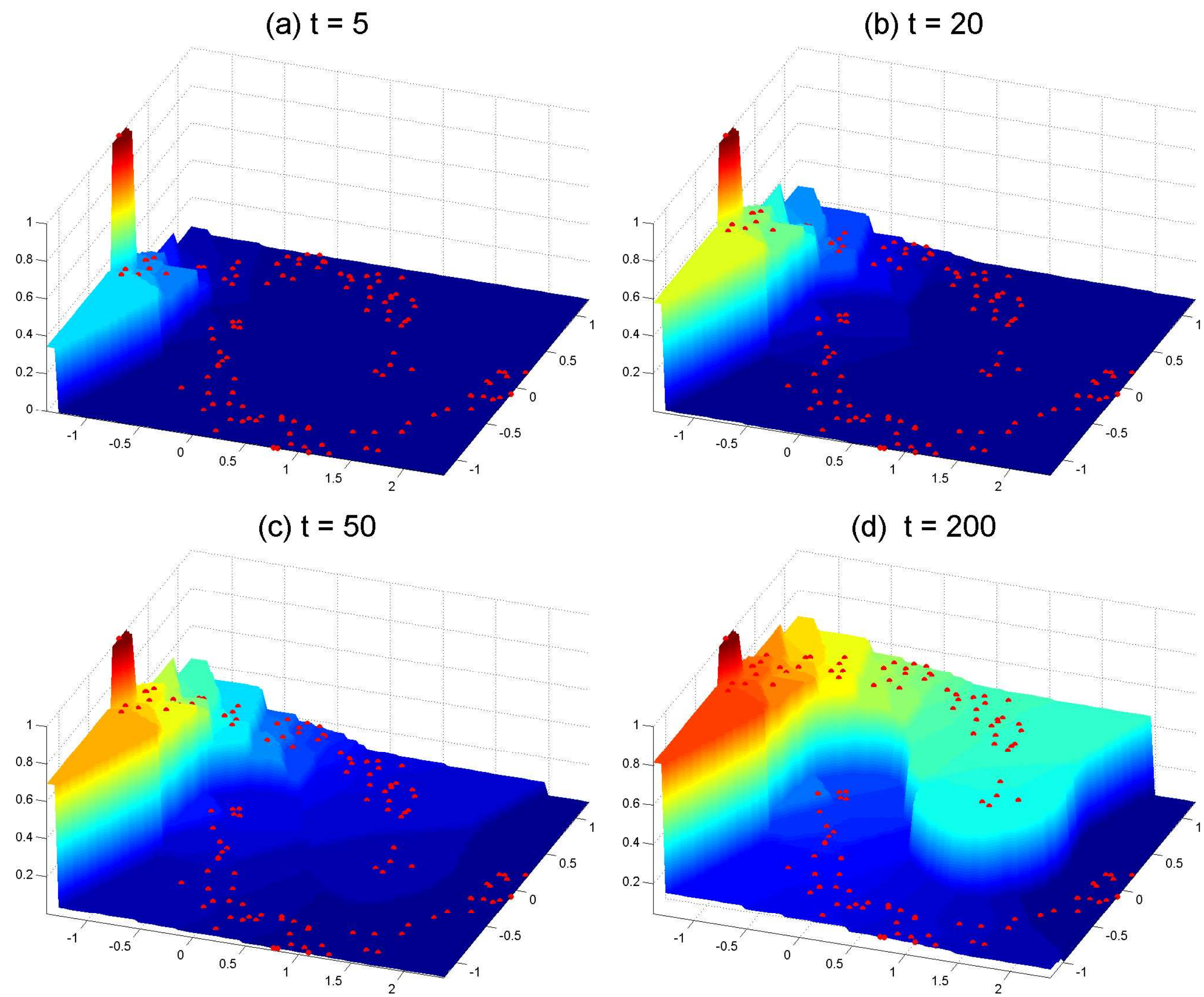
- The relevant degrees of points in the upper moon to the query should decrease along the moon shape.
- All points in the upper moon should be more relevant to the query than the points in the lower moon.

Toy Ranking



- Simply ranking the data according to the shortest paths on the graph does not work well.
- Robust solution is to assemble all paths between two points: $f^* = \sum_i \alpha^i S^i y$.

Toy Ranking



Connection to Google

Theorem. *For the task of ranking data represented by a connected and undirected graph without queries, f^* and PageRank yield the same ranking list.*

Personalized Google: a variant

The ranking scores given by PageRank:

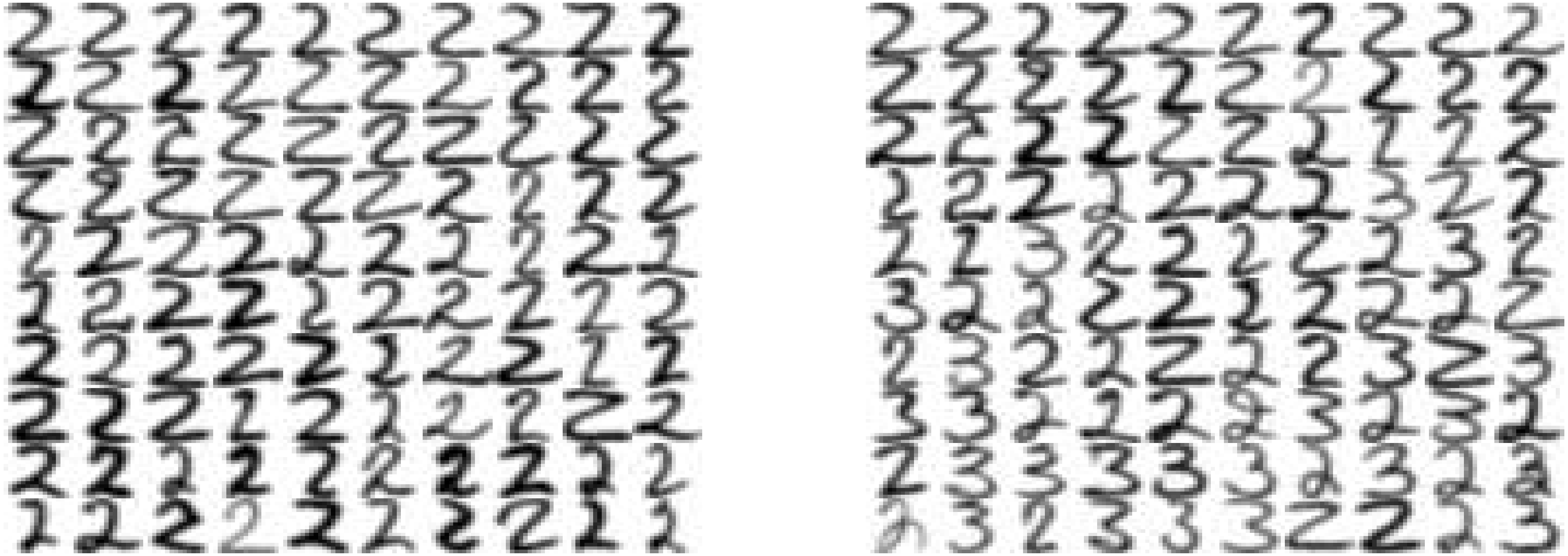
$$\pi(t+1) = \alpha P^T \pi(t). \tag{4}$$

Add a query term on the right-hand side for the query-based ranking,

$$\pi(t+1) = \alpha P^T \pi(t) + (1-\alpha) y. \tag{5}$$

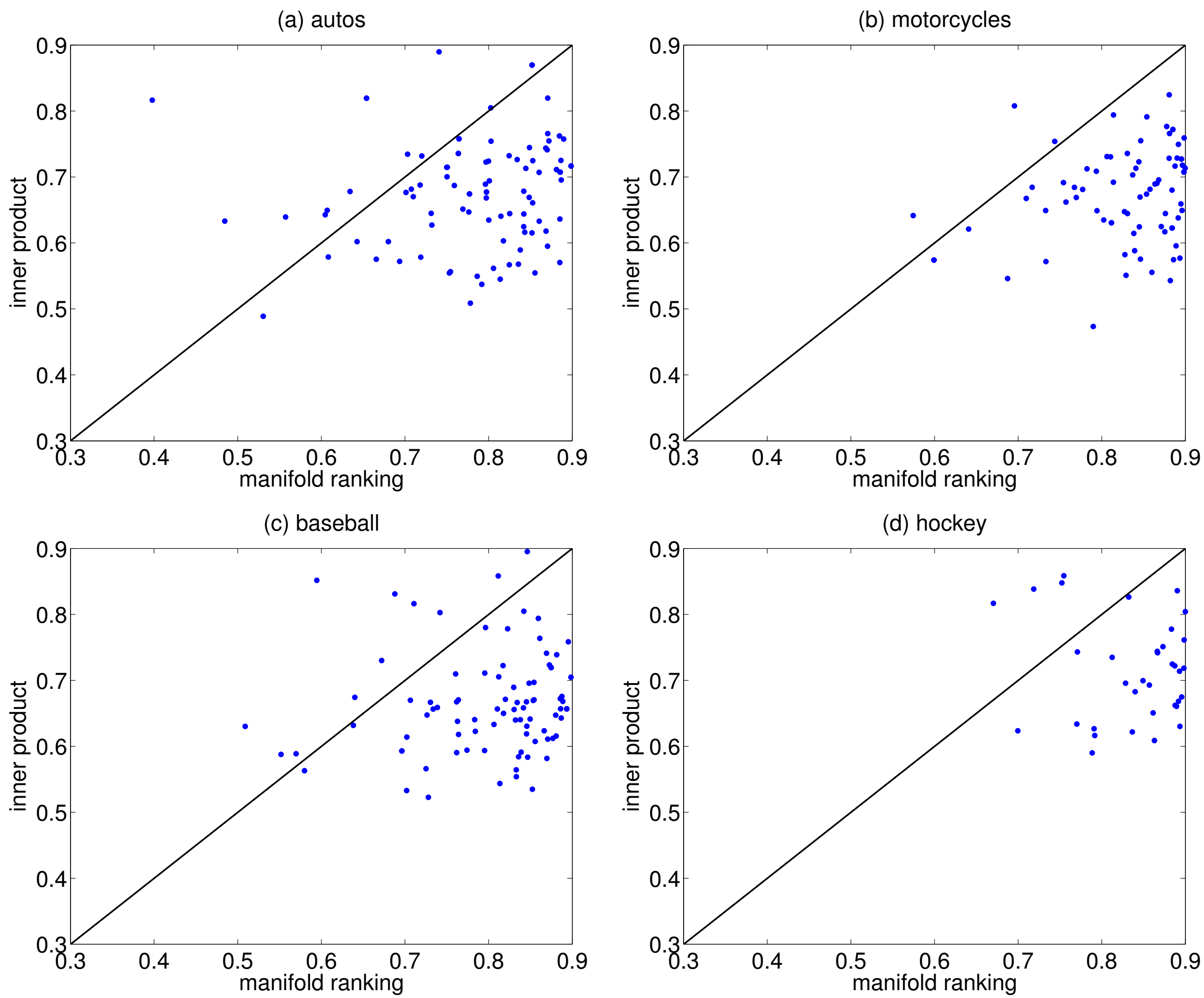
This can be viewed as the *personalized* version of PageRank.

Image Ranking



The top-left digit in each panel is the query. The left panel shows the top 99 by our method; and the right panel shows the top 99 by the Euclidean distance.

Document Ranking



Related Work

- Graph/disffusion/cluster kernel (Kondor et al 2002; Chapelle et al. 2002; Smola et al. 2003).
- Spectral clustering (Shi et al. 1997; Ng et al. 2001).
- Manifold learning (nonlinear data reduction)(Tenenbaum et al. 2000; Roweis et al. 2000)

Related Work

- Random walks (Szummer et al. 2001).
- Graph min-cuts (Blum et al. 2001)
- Learning on manifolds (Belkin et al. 2001).
- Gaussian random fields (Zhu et al. 2003).

Conclusion

- Proposed a general semi-supervised learning algorithm.
- Proposed a general example-based ranking algorithm.

Next Work

- Model selection.
- Active learning.
- Generalization theory of learning from labeled and unlabeled data.
- Specifical problems & large-scale problems.

References

1. Zhou, D., Bousquet, O., Lal, T.N., Weston, J. and Schölkopf, B.: **Learning with Local and Global Consistency**. NIPS, 2003.
2. Zhou, D., Weston, J., Gretton, A., Bousquet, O. and Schölkopf, B.: **Ranking on Data Manifolds**. NIPS, 2003.
3. Weston, J., C. Leslie, D. Zhou, A. Elisseeff and W. S. Noble: **Semi-Supervised Protein Classification using Cluster Kernels**. NIPS, 2003.