# Discriminative Experimental Design

**Yu Zhang** and Dit-Yan Yeung

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology

ECML PKDD 2011

# Outline

# Active Learning

# Active Learning

- Active learning selects unlabeled data points to query some oracle.

# Active Learning

- Active learning selects unlabeled data points to query some oracle.
- Existing active learning methods: uncertainty sampling (SVM Active Learning), query-by-committee, representative sampling (Transductive Experimental Design).

# Active Learning

- Active learning selects unlabeled data points to query some oracle.
- Existing active learning methods: uncertainty sampling (SVM Active Learning), query-by-committee, representative sampling (Transductive Experimental Design).

**Input:** Labeled data set $\mathcal{L}$; Unlabeled data set $\mathcal{U}$
**Output:** Learning model

Step 1: Train a learning model based on $\mathcal{L}$;
Step 2:
For $t = 1, \ldots, t_{\text{max}}$
2.1: Select an unlabeled data set $\mathcal{S}$ from $\mathcal{U}$ based on some unlabeled data selection criterion;
2.2: Query an oracle to label $\mathcal{S}$;
2.3: $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{S}, \mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{S}$;
2.4: Re-train the learning model based on $\mathcal{L}$;

# Our Contribution

# Our Contribution

- Some methods are complementary.

# Our Contribution

- Some methods are complementary.
  - SVM Active Learning: Use of discriminative information; Selection of one point in an iteration

# Our Contribution

- Some methods are complementary.
    - SVM Active Learning: Use of discriminative information; Selection of one point in an iteration
    - TED: Use of data distribution information; Selection of multiple points in an iteration

# Our Contribution

- Some methods are complementary.
  - SVM Active Learning: Use of discriminative information; Selection of one point in an iteration
  - TED: Use of data distribution information; Selection of multiple points in an iteration

- Our Contributions:

# Our Contribution

- Some methods are complementary.
  - SVM Active Learning: Use of discriminative information; Selection of one point in an iteration
  - TED: Use of data distribution information; Selection of multiple points in an iteration

- Our Contributions:
  - The proposal of discriminative experimental design (DED), combining the strengths of both SVM active learning and TED.

# Our Contribution

- Some methods are complementary.
  - SVM Active Learning: Use of discriminative information; Selection of one point in an iteration
  - TED: Use of data distribution information; Selection of multiple points in an iteration
- Our Contributions:
  - The proposal of discriminative experimental design (DED), combining the strengths of both SVM active learning and TED.
  - A projection method to solve the optimization problem.

# Our Contribution

- Some methods are complementary.
  - SVM Active Learning: Use of discriminative information; Selection of one point in an iteration
  - TED: Use of data distribution information; Selection of multiple points in an iteration
- Our Contributions:
  - The proposal of discriminative experimental design (DED), combining the strengths of both SVM active learning and TED.
  - A projection method to solve the optimization problem.
  - The good performance on some benchmark datasets.

# Outline

# Notations

# Notations

- $\mathbf{V} \in \mathbb{R}^{d \times n}$: The matrix for the unlabeled data currently available

# Notations

- $\mathbf{V} \in \mathbb{R}^{d \times n}$: The matrix for the unlabeled data currently available
- $\mathbf{X} \in \mathbb{R}^{d \times t}$: The selected subset of unlabeled data

# Notations

- $\mathbf{V} \in \mathbb{R}^{d \times n}$: The matrix for the unlabeled data currently available
- $\mathbf{X} \in \mathbb{R}^{d \times t}$: The selected subset of unlabeled data
- $t$: The number of selected data points

# Notations

- $\mathbf{V} \in \mathbb{R}^{d \times n}$: The matrix for the unlabeled data currently available
- $\mathbf{X} \in \mathbb{R}^{d \times t}$: The selected subset of unlabeled data
- $t$: The number of selected data points
- $\phi(\cdot)$: The feature mapping corresponding to some kernel function $k(\cdot, \cdot)$

# Outline

# Least-Square SVM Revisited

# Least-Square SVM Revisited

- The objective function of least-square SVM is formulated as:

$$\min_{\mathbf{w}} \sum_{i=1}^{l} (\mathbf{w}^T \phi(\mathbf{x}_i) - y_i)^2 + \lambda \|\mathbf{w}\|_2^2. \quad (1)$$

# Least-Square SVM Revisited

- The objective function of least-square SVM is formulated as:

$$\min_{\mathbf{w}} \sum_{i=1}^{l} (\mathbf{w}^T \phi(\mathbf{x}_i) - y_i)^2 + \lambda \|\mathbf{w}\|_2^2. \quad (1)$$

- Its equivalent form:

$$\min_{\mathbf{w}} J(\mathbf{w}) = \sum_{i=1}^{l} (1 - y_i \mathbf{w}^T \phi(\mathbf{x}_i))^2 + \lambda \|\mathbf{w}\|_2^2. \quad (2)$$

# Least-Square SVM Revisited

- The objective function of least-square SVM is formulated as:

$$\min_{\mathbf{w}} \sum_{i=1}^{l} (\mathbf{w}^T \phi(\mathbf{x}_i) - y_i)^2 + \lambda \|\mathbf{w}\|_2^2. \quad (1)$$

- Its equivalent form:

$$\min_{\mathbf{w}} J(\mathbf{w}) = \sum_{i=1}^{l} (1 - y_i \mathbf{w}^T \phi(\mathbf{x}_i))^2 + \lambda \|\mathbf{w}\|_2^2. \quad (2)$$

- Here the square loss is similar to the square hinge loss $L'(s,t) = \max(0, 1 - st)^2$.

# Least-Square SVM Revisited

- The objective function of least-square SVM is formulated as:

$$\min_{\mathbf{w}} \sum_{i=1}^{l} (\mathbf{w}^T \phi(\mathbf{x}_i) - y_i)^2 + \lambda \|\mathbf{w}\|_2^2. \quad (1)$$

- Its equivalent form:

$$\min_{\mathbf{w}} J(\mathbf{w}) = \sum_{i=1}^{l} (1 - y_i \mathbf{w}^T \phi(\mathbf{x}_i))^2 + \lambda \|\mathbf{w}\|_2^2. \quad (2)$$

- Here the square loss is similar to the square hinge loss $L'(s,t) = \max(0, 1 - st)^2$.
- The function score for a data point is defined as:

$$y = \frac{1}{\mathbf{w}^T \phi(\mathbf{x})},$$

# The Objective Function

# The Objective Function

- According to the analysis in TED, the estimation error satisfies

$$\mathrm{cov}(\mathbf{w} - \mathbf{w}^\star) \propto \mathbf{C}_\mathbf{w} = \left(\frac{\partial^2 J(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^T}\right)^{-1} = \left(\phi(\mathbf{X})\mathbf{Y}_\mathbf{X}^2\phi(\mathbf{X})^T + \lambda \mathbf{I}_{d'}\right)^{-1}$$

# The Objective Function

- According to the analysis in TED, the estimation error satisfies

$$\mathrm{cov}(\mathbf{w}-\mathbf{w}^{\star}) \propto \mathbf{C_w} = \left(\frac{\partial^2 J(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^T}\right)^{-1} = \left(\phi(\mathbf{X})\mathbf{Y}_{\mathbf{X}}^2\phi(\mathbf{X})^T + \lambda \mathbf{I}_{d'}\right)^{-1}$$

- The predictive error on the whole unlabeled data set satisfies

$$\mathbf{C_f} = \mathbf{Y_V}\phi(\mathbf{V})^T\mathbf{C_w}\phi(\mathbf{V})\mathbf{Y_V}$$

# The Objective Function

- According to the analysis in TED, the estimation error satisfies

$$\mathrm{cov}(\mathbf{w} - \mathbf{w}^\star) \propto \mathbf{C_w} = \left(\frac{\partial^2 J(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^T}\right)^{-1} = \left(\phi(\mathbf{X})\mathbf{Y}_\mathbf{X}^2 \phi(\mathbf{X})^T + \lambda \mathbf{I}_{d'}\right)^{-1}$$

- The predictive error on the whole unlabeled data set satisfies

$$\mathbf{C_f} = \mathbf{Y_V}\phi(\mathbf{V})^T \mathbf{C_w} \phi(\mathbf{V})\mathbf{Y_V}$$

- The A-optimal design is used to minimize the predictive variance:

$$\min \mathrm{tr}(\mathbf{C_f}).$$

# The Objective Function

- According to the analysis in TED, the estimation error satisfies

$$\operatorname{cov}(\mathbf{w}-\mathbf{w}^{\star}) \propto \mathbf{C}_{\mathbf{w}} = \left(\frac{\partial^2 J(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^T}\right)^{-1} = \left(\phi(\mathbf{X})\mathbf{Y}_{\mathbf{X}}^2\phi(\mathbf{X})^T + \lambda \mathbf{I}_{d'}\right)^{-1}$$

- The predictive error on the whole unlabeled data set satisfies

$$\mathbf{C}_{\mathbf{f}} = \mathbf{Y}_{\mathbf{V}}\phi(\mathbf{V})^T\mathbf{C}_{\mathbf{w}}\phi(\mathbf{V})\mathbf{Y}_{\mathbf{V}}$$

- The A-optimal design is used to minimize the predictive variance:

$$\min \operatorname{tr}(\mathbf{C}_{\mathbf{f}}).$$

Definition

Discriminative Experimental Design:

$$\begin{aligned} \max_{\mathbf{X},\mathbf{Y}_{\mathbf{X}}} \quad & \operatorname{tr}\left[\mathbf{Y}_{\mathbf{V}}\mathbf{K}_{\mathbf{VX}}\mathbf{Y}_{\mathbf{X}}(\lambda \mathbf{I}_t + \mathbf{Y}_{\mathbf{X}}\mathbf{K}_{\mathbf{X}}\mathbf{Y}_{\mathbf{X}})^{-1}\mathbf{Y}_{\mathbf{X}}\mathbf{K}_{\mathbf{XV}}\mathbf{Y}_{\mathbf{V}}\right] \\ \text{s.t.} \quad & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = t, \mathbf{Y}_{\mathbf{X}} \subset \mathbf{Y}_{\mathbf{V}}. \end{aligned} \tag{3}$$

# The Relationship between DED and TED

# The Relationship between DED and TED

- The optimization problem of linear DED:

$$\begin{aligned} \max_{\tilde{\mathbf{X}}} \quad & \mathrm{tr}\left[\tilde{\mathbf{V}}^T\tilde{\mathbf{X}}(\lambda\mathbf{I}_t + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\right] \\ \text{s.t.} \quad & \tilde{\mathbf{X}} \subset \tilde{\mathbf{V}}, |\tilde{\mathbf{X}}| = t. \end{aligned} \tag{4}$$

# The Relationship between DED and TED

- The optimization problem of linear DED:

$$\begin{aligned} \max_{\tilde{\mathbf{X}}} \quad & \mathrm{tr}\left[\tilde{\mathbf{V}}^T\tilde{\mathbf{X}}(\lambda\mathbf{I}_t + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\right] \\ \text{s.t.} \quad & \tilde{\mathbf{X}} \subset \tilde{\mathbf{V}}, |\tilde{\mathbf{X}}| = t. \end{aligned} \tag{4}$$

- This is identical to the optimization problem of TED.

# The Relationship between DED and TED

- The optimization problem of linear DED:

$$\begin{aligned} \max_{\tilde{\mathbf{X}}} \quad & \mathrm{tr}\left[\tilde{\mathbf{V}}^T\tilde{\mathbf{X}}(\lambda\mathbf{I}_t + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\right] \\ \text{s.t.} \quad & \tilde{\mathbf{X}} \subset \tilde{\mathbf{V}}, |\tilde{\mathbf{X}}| = t. \end{aligned} \tag{4}$$

- This is identical to the optimization problem of TED.
- TED can be seen as a special case of DED.

# The Relationship between DED and TED

- The optimization problem of linear DED:

$$\begin{aligned} \max_{\tilde{\mathbf{X}}} \quad & \mathrm{tr}\left[\tilde{\mathbf{V}}^T\tilde{\mathbf{X}}(\lambda\mathbf{I}_t + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\right] \\ \text{s.t.} \quad & \tilde{\mathbf{X}} \subset \tilde{\mathbf{V}}, |\tilde{\mathbf{X}}| = t. \end{aligned} \tag{4}$$

- This is identical to the optimization problem of TED.
- TED can be seen as a special case of DED.
- DED is a weighted version of TED.

# The Relationship between DED and TED

- The optimization problem of linear DED:

$$\begin{aligned} \max_{\tilde{\mathbf{X}}} \quad & \mathrm{tr}\left[\tilde{\mathbf{V}}^T\tilde{\mathbf{X}}(\lambda\mathbf{I}_t + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\right] \\ \text{s.t.} \quad & \tilde{\mathbf{X}} \subset \tilde{\mathbf{V}}, |\tilde{\mathbf{X}}| = t. \end{aligned} \tag{4}$$

- This is identical to the optimization problem of TED.
- TED can be seen as a special case of DED.
- DED is a weighted version of TED.
  - The weights are related to function scores of the data points.

# Reformulation of DED

# Reformulation of DED

- A selection indicator matrix $\mathbf{S} \in \{0,1\}^{n \times t}$ is defined as

$$s_{ij} = \begin{cases} 1 & \text{if } (\phi(\mathbf{X})\mathbf{Y_X})_{,j} \text{ is from } (\phi(\mathbf{V})\mathbf{Y_V})_{,i} \\ 0 & \text{otherwise} \end{cases}$$

# Reformulation of DED

- A selection indicator matrix $\mathbf{S} \in \{0, 1\}^{n \times t}$ is defined as

$$s_{ij} = \begin{cases} 1 & \text{if } (\phi(\mathbf{X})\mathbf{Y}_{\mathbf{X}})_{,j} \text{ is from } (\phi(\mathbf{V})\mathbf{Y}_{\mathbf{V}})_{,i} \\ 0 & \text{otherwise} \end{cases}$$

- The constraint set for $\mathbf{S}$ is $C_S = \{\mathbf{S} \,|\, \mathbf{S} \in \{0, 1\}^{n \times t}, \mathbf{S}^T\mathbf{S} = \mathbf{I}_t\}$.

# Reformulation of DED

- A selection indicator matrix $\mathbf{S} \in \{0,1\}^{n\times t}$ is defined as

$$s_{ij} = \begin{cases} 1 & \text{if } (\phi(\mathbf{X})\mathbf{Y_X})_{,j} \text{ is from } (\phi(\mathbf{V})\mathbf{Y_V})_{,i} \\ 0 & \text{otherwise} \end{cases}$$

- The constraint set for $\mathbf{S}$ is $C_S = \left\{\mathbf{S} \,|\, \mathbf{S} \in \{0,1\}^{n\times t}, \mathbf{S}^T\mathbf{S} = \mathbf{I}_t\right\}$.
- The objective function of DED can be reformulated as

$$\begin{aligned} \max_{\mathbf{S}} \quad & \mathrm{tr}\left[(\mathbf{S}^T(\lambda\mathbf{I}_n + \tilde{\mathbf{K}}_\mathbf{V})\mathbf{S})^{-1}\mathbf{S}^T\tilde{\mathbf{K}}_\mathbf{V}^2\mathbf{S}\right] \\ \text{s.t.} \quad & \mathbf{S} \in C_S, \end{aligned} \tag{5}$$

# Reformulation of DED

- A selection indicator matrix $\mathbf{S} \in \{0,1\}^{n \times t}$ is defined as

$$s_{ij} = \begin{cases} 1 & \text{if } (\phi(\mathbf{X})\mathbf{Y}_{\mathbf{X}})_{,j} \text{ is from } (\phi(\mathbf{V})\mathbf{Y}_{\mathbf{V}})_{,i} \\ 0 & \text{otherwise} \end{cases}$$

- The constraint set for $\mathbf{S}$ is $C_S = \left\{\mathbf{S} \,|\, \mathbf{S} \in \{0,1\}^{n \times t}, \mathbf{S}^T\mathbf{S} = \mathbf{I}_t\right\}$.
- The objective function of DED can be reformulated as

$$\begin{aligned} \max_{\mathbf{S}} \quad & \mathrm{tr}\left[(\mathbf{S}^T(\lambda \mathbf{I}_n + \tilde{\mathbf{K}}_{\mathbf{V}})\mathbf{S})^{-1}\mathbf{S}^T\tilde{\mathbf{K}}_{\mathbf{V}}^2\mathbf{S}\right] \\ \text{s.t.} \quad & \mathbf{S} \in C_S, \end{aligned} \tag{5}$$

- If there is no constraint, the optimal solution has the form of $\mathbf{S}^\star\mathbf{P}$.

# Reformulation of DED

- A selection indicator matrix $\mathbf{S} \in \{0,1\}^{n\times t}$ is defined as

$$s_{ij} = \begin{cases} 1 & \text{if } (\phi(\mathbf{X})\mathbf{Y}_\mathbf{X})_{,j} \text{ is from } (\phi(\mathbf{V})\mathbf{Y}_\mathbf{V})_{,i} \\ 0 & \text{otherwise} \end{cases}$$

- The constraint set for $\mathbf{S}$ is $C_S = \{\mathbf{S} \,|\, \mathbf{S} \in \{0,1\}^{n\times t}, \mathbf{S}^T\mathbf{S} = \mathbf{I}_t\}$.
- The objective function of DED can be reformulated as

$$\begin{aligned} \max_{\mathbf{S}} \quad & \mathrm{tr}\left[(\mathbf{S}^T(\lambda\mathbf{I}_n + \tilde{\mathbf{K}}_\mathbf{V})\mathbf{S})^{-1}\mathbf{S}^T\tilde{\mathbf{K}}_\mathbf{V}^2\mathbf{S}\right] \\ \text{s.t.} \quad & \mathbf{S} \in C_S, \end{aligned} \qquad (5)$$

- If there is no constraint, the optimal solution has the form of $\mathbf{S}^\star\mathbf{P}$.
  - $\mathbf{S}^\star$ consists of the top $t$ eigenvectors of $\tilde{\mathbf{K}}_\mathbf{V}$.

# Reformulation of DED

- A selection indicator matrix $\mathbf{S} \in \{0,1\}^{n\times t}$ is defined as

$$s_{ij} = \begin{cases} 1 & \text{if } (\phi(\mathbf{X})\mathbf{Y}_{\mathbf{X}})_{,j} \text{ is from } (\phi(\mathbf{V})\mathbf{Y}_{\mathbf{V}})_{,i} \\ 0 & \text{otherwise} \end{cases}$$

- The constraint set for $\mathbf{S}$ is $C_S = \left\{\mathbf{S} \,|\, \mathbf{S} \in \{0,1\}^{n\times t}, \mathbf{S}^T\mathbf{S} = \mathbf{I}_t\right\}$.
- The objective function of DED can be reformulated as

$$\begin{aligned} \max_{\mathbf{S}} \quad & \mathrm{tr}\left[(\mathbf{S}^T(\lambda\mathbf{I}_n + \tilde{\mathbf{K}}_{\mathbf{V}})\mathbf{S})^{-1}\mathbf{S}^T\tilde{\mathbf{K}}_{\mathbf{V}}^2\mathbf{S}\right] \\ \text{s.t.} \quad & \mathbf{S} \in C_S, \end{aligned} \tag{5}$$

- If there is no constraint, the optimal solution has the form of $\mathbf{S}^\star\mathbf{P}$.
  - $\mathbf{S}^\star$ consists of the top $t$ eigenvectors of $\tilde{\mathbf{K}}_{\mathbf{V}}$.
  - $\mathbf{P} \in \mathbb{R}^{t\times t}$ is an orthogonal matrix.

# The Projection Method

# The Projection Method

- The optimal solution $\mathbf{S}^{\star}\mathbf{P}$ is projected to the set $C_S$:

$$\min_{\mathbf{P},\mathbf{Q}} \quad \|\mathbf{S}^{\star}\mathbf{P} - \mathbf{Q}\|_F^2$$
$$\text{s.t.} \quad \mathbf{Q} \in C_S, \ \mathbf{P}\mathbf{P}^T = \mathbf{I}_t, \qquad (6)$$

# The Projection Method

- The optimal solution $\mathbf{S}^\star\mathbf{P}$ is projected to the set $C_S$:

$$\min_{\mathbf{P},\mathbf{Q}} \quad \|\mathbf{S}^\star\mathbf{P}-\mathbf{Q}\|_F^2 \qquad \text{s.t.} \quad \mathbf{Q}\in C_S,\ \mathbf{P}\mathbf{P}^T=\mathbf{I}_t, \tag{6}$$

- Its equivalent form:

$$\max_{\mathbf{P},\mathbf{Q}} \quad \mathrm{tr}(\mathbf{Q}^T\mathbf{S}^\star\mathbf{P}) \qquad \text{s.t.} \quad \mathbf{Q}\in C_S,\ \mathbf{P}\mathbf{P}^T=\mathbf{I}_t. \tag{7}$$

# The Projection Method

- The optimal solution $\mathbf{S}^{\star}\mathbf{P}$ is projected to the set $C_S$:

$$\min_{\mathbf{P},\mathbf{Q}} \quad \|\mathbf{S}^{\star}\mathbf{P} - \mathbf{Q}\|_F^2 \qquad \text{s.t.} \quad \mathbf{Q} \in C_S,\ \mathbf{P}\mathbf{P}^T = \mathbf{I}_t, \tag{6}$$

- Its equivalent form:

$$\max_{\mathbf{P},\mathbf{Q}} \quad \mathrm{tr}(\mathbf{Q}^T\mathbf{S}^{\star}\mathbf{P}) \qquad \text{s.t.} \quad \mathbf{Q} \in C_S,\ \mathbf{P}\mathbf{P}^T = \mathbf{I}_t. \tag{7}$$

- An alternating optimization method is used to solve this problem.

# Subproblem 1

# Subproblem 1

- When **P** is fixed, the optimization problem with respect to **Q** is

$$\max_{\mathbf{Q}} \operatorname{tr}(\mathbf{Q}^T \mathbf{S}^\star \mathbf{P})$$
$$\text{s.t. } \mathbf{Q} \in \{0,1\}^{n\times t}, \mathbf{Q}^T \mathbf{1}_n = \mathbf{1}_t, \mathbf{Q}\mathbf{1}_t \leq \mathbf{1}_n. \quad (8)$$

# Subproblem 1

- When **P** is fixed, the optimization problem with respect to **Q** is

$$\max_{\mathbf{Q}} \operatorname{tr}(\mathbf{Q}^T \mathbf{S}^\star \mathbf{P})$$
$$\text{s.t. } \mathbf{Q} \in \{0,1\}^{n\times t}, \mathbf{Q}^T \mathbf{1}_n = \mathbf{1}_t, \mathbf{Q}\mathbf{1}_t \le \mathbf{1}_n. \quad (8)$$

- This is an integer programming problem with no efficient solution.

# Subproblem 1

- When **P** is fixed, the optimization problem with respect to **Q** is

$$\max_{\mathbf{Q}} \operatorname{tr}(\mathbf{Q}^T \mathbf{S}^\star \mathbf{P})$$
$$\text{s.t. } \mathbf{Q} \in \{0, 1\}^{n \times t}, \mathbf{Q}^T \mathbf{1}_n = \mathbf{1}_t, \mathbf{Q}\mathbf{1}_t \leq \mathbf{1}_n. \quad (8)$$

- This is an integer programming problem with no efficient solution.
- This problem is to find the $t$ largest elements in $\mathbf{S}^\star\mathbf{P}$

# Subproblem 1

- When $\mathbf{P}$ is fixed, the optimization problem with respect to $\mathbf{Q}$ is

$$\max_{\mathbf{Q}} \operatorname{tr}(\mathbf{Q}^T\mathbf{S}^\star\mathbf{P})$$
$$\text{s.t. } \mathbf{Q} \in \{0,1\}^{n\times t}, \mathbf{Q}^T\mathbf{1}_n = \mathbf{1}_t, \mathbf{Q}\mathbf{1}_t \leq \mathbf{1}_n. \quad (8)$$

- This is an integer programming problem with no efficient solution.
- This problem is to find the $t$ largest elements in $\mathbf{S}^\star\mathbf{P}$
  - No two elements can be in the same column or the same row.

# Subproblem 1

- When $\mathbf{P}$ is fixed, the optimization problem with respect to $\mathbf{Q}$ is

$$\max_{\mathbf{Q}} \operatorname{tr}(\mathbf{Q}^T \mathbf{S}^\star \mathbf{P})$$
$$\text{s.t. } \mathbf{Q} \in \{0,1\}^{n\times t}, \mathbf{Q}^T \mathbf{1}_n = \mathbf{1}_t, \mathbf{Q}\mathbf{1}_t \leq \mathbf{1}_n. \quad (8)$$

- This is an integer programming problem with no efficient solution.
- This problem is to find the $t$ largest elements in $\mathbf{S}^\star\mathbf{P}$
  - No two elements can be in the same column or the same row.
- Observation: the largest elements of different columns in $\mathbf{S}^\star\mathbf{P}$ usually lie in different rows.

# Subproblem 1

- When $\mathbf{P}$ is fixed, the optimization problem with respect to $\mathbf{Q}$ is

$$\max_{\mathbf{Q}} \operatorname{tr}(\mathbf{Q}^T \mathbf{S}^\star \mathbf{P})$$
$$\text{s.t. } \mathbf{Q} \in \{0, 1\}^{n \times t}, \mathbf{Q}^T \mathbf{1}_n = \mathbf{1}_t, \mathbf{Q}\mathbf{1}_t \leq \mathbf{1}_n. \quad (8)$$

- This is an integer programming problem with no efficient solution.
- This problem is to find the $t$ largest elements in $\mathbf{S}^\star\mathbf{P}$
  - No two elements can be in the same column or the same row.
- Observation: the largest elements of different columns in $\mathbf{S}^\star\mathbf{P}$ usually lie in different rows.
- We propose a greedy method to select multiple largest elements in different rows.

# Subproblem 2

# Subproblem 2

- When $\mathbf{Q}$ is fixed, the optimization problem with respect to $\mathbf{P}$ is

$$\begin{aligned} \max_{\mathbf{P}} \quad & \mathrm{tr}(\mathbf{Q}^T\mathbf{S}^{\star}\mathbf{P}) \\ \text{s.t.} \quad & \mathbf{P}\mathbf{P}^T = \mathbf{I}_t. \end{aligned} \tag{9}$$

# Subproblem 2

- When $\mathbf{Q}$ is fixed, the optimization problem with respect to $\mathbf{P}$ is

$$\begin{aligned} \max_{\mathbf{P}} \quad & \mathrm{tr}(\mathbf{Q}^T \mathbf{S}^\star \mathbf{P}) \\ \text{s.t.} \quad & \mathbf{P}\mathbf{P}^T = \mathbf{I}_t. \end{aligned} \tag{9}$$

- By using Lagrangian multiplier method, we can get the analytical solution as

$$\mathbf{P}^\star = \mathbf{U}\mathbf{R}^T.$$

# Subproblem 2

- When $\mathbf{Q}$ is fixed, the optimization problem with respect to $\mathbf{P}$ is

$$\max_{\mathbf{P}} \quad \mathrm{tr}(\mathbf{Q}^T\mathbf{S}^\star\mathbf{P})$$
$$\text{s.t.} \quad \mathbf{P}\mathbf{P}^T = \mathbf{I}_t. \tag{9}$$

- By using Lagrangian multiplier method, we can get the analytical solution as

$$\mathbf{P}^\star = \mathbf{U}\mathbf{R}^T.$$

  - $(\mathbf{S}^\star)^T\mathbf{Q} = \mathbf{U}\Sigma\mathbf{R}^T$ be the singular value decomposition.

# Properties of Our Optimization Method

# Properties of Our Optimization Method

- The computational complexity of our method is $O(n^2 t)$.

# Properties of Our Optimization Method

- The computational complexity of our method is $O(n^2t)$.
- DED is insensitive to the regularization parameter.

# Outline

# Experimental Setup

# Experimental Setup

- The method compared: DED, SVM active learning, TED, batch mode active learning.

# Experimental Setup

- The method compared: DED, SVM active learning, TED, batch mode active learning.
- Two public benchmark data sets used: Newsgroups and Reuters.

# Experimental Setup

- The method compared: DED, SVM active learning, TED, batch mode active learning.
- Two public benchmark data sets used: Newsgroups and Reuters.
- Performance measure: The area under the ROC curve (AUC).

# Experimental Setup

- The method compared: DED, SVM active learning, TED, batch mode active learning.
- Two public benchmark data sets used: Newsgroups and Reuters.
- Performance measure: The area under the ROC curve (AUC).
- The size of queries $t$: 5

# Experimental Setup

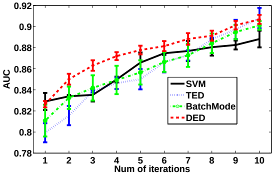
- The method compared: DED, SVM active learning, TED, batch mode active learning.
- Two public benchmark data sets used: Newsgroups and Reuters.
- Performance measure: The area under the ROC curve (AUC).
- The size of queries $t$: 5
- The regularization parameters: 0.01.

# Experimental Setup

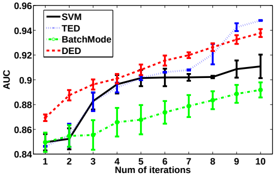
- The method compared: DED, SVM active learning, TED, batch mode active learning.
- Two public benchmark data sets used: Newsgroups and Reuters.
- Performance measure: The area under the ROC curve (AUC).
- The size of queries $t$: 5
- The regularization parameters: 0.01.
- Five labeled data points are provided for each class before active learning starts.

# Results on Newsgroups Data

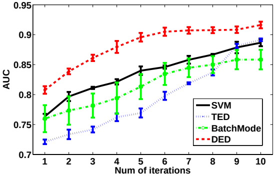
# Results on Newsgroups Data



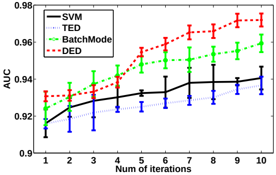

(e) Autos




(f) Motorcycles

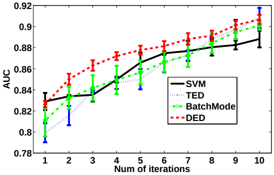



(g) Baseball

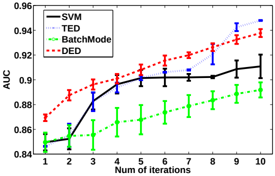



(h) Hockey

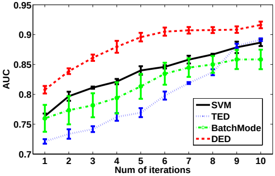
# Results on Newsgroups Data



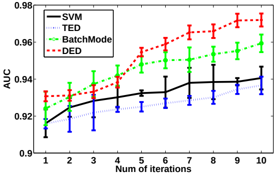

(i) Autos




(j) Motorcycles




(k) Baseball

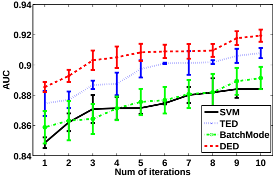



(l) Hockey

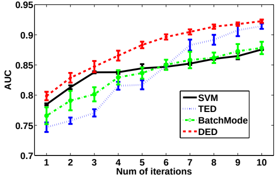
- When the labeled data is scarce, data distribution information is very important.

# Results on Reuters Data

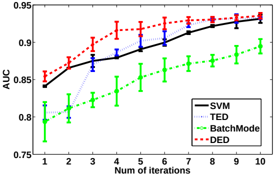
# Results on Reuters Data



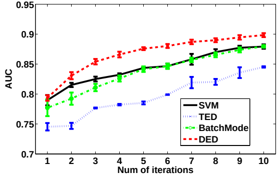

(q) CCAT




(r) ECAT



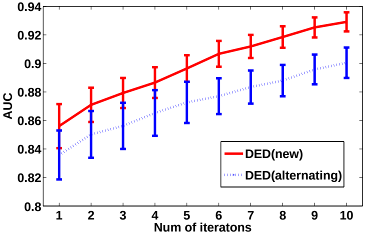

(s) GCAT



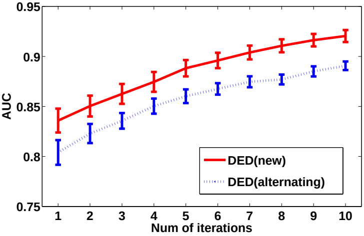

(t) MCAT

# Comparison on Two Optimization Techniques

# Comparison on Two Optimization Techniques




(w) Newsgroups data




(x) Reuters data

# Outline

# Conclusion

# Conclusion

- A novel active learning method has been proposed.

# Conclusion

- A novel active learning method has been proposed.
- The data selection criterion utilizes discriminative information and data distribution information.

# Conclusion

- A novel active learning method has been proposed.
- The data selection criterion utilizes discriminative information and data distribution information.

Future Work:

# Conclusion

- A novel active learning method has been proposed.
- The data selection criterion utilizes discriminative information and data distribution information.

Future Work:

- The integration of active learning and semi-supervised learning

Thanks very much for your attention!