Recap

- Problem: more data ≠ higher performance
- Goal: select small informative subset from a large dataset
- Approach:
  - Representative: Core-Set, Clustering
  - Inconsistency: QBC
  - Gradient-based
- **Diversity**
  - ➢Determinantal Point Processes (DPP)
  - ➢Transductive Experimental Design (TED)

Determinantal Point Processes for Mini-Batch Diversification

Cheng Zhang
Disney Research
Pittsburgh, PA, USA
cheng.zhang@disneyresearch.com

Hedvig Kjellström
KTH Royal Institute of Technology
Stockholm, Sweden
hedvig@kth.se

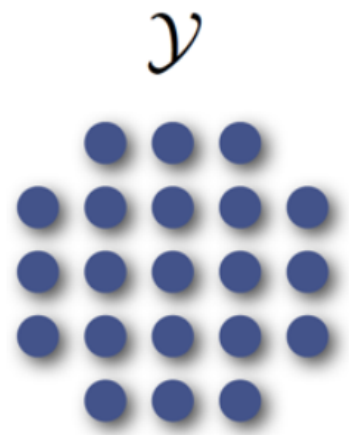
Stephan Mandt
Disney Research
Pittsburgh, PA, USA
stephan.mandt@disneyresearch.com

UAI, 2017

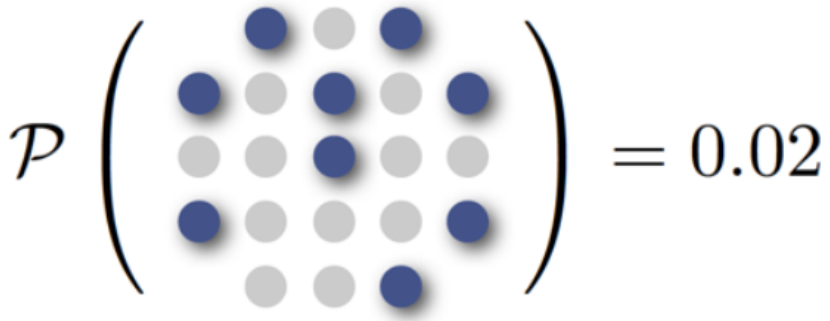
Introduction

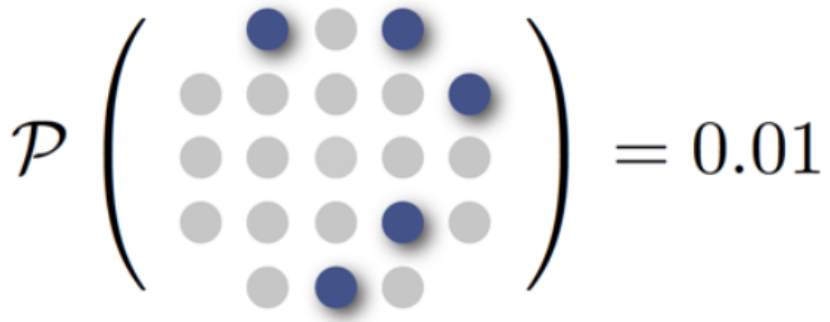
- Real-world data sets are naturally imbalanced, e.g.
  - sports topic appears more often in the news than biology
  - Internet contains more images of young people than of senior people
  - Youtube has more videos of cats than of bees or ants
- A biased mini-batch subsampling scheme for imbalanced data with Determinantal Point Processes (DPP).

Discrete point process



Ground set





Prob. of observing a subset

Discrete point process

- N items (e.g., images or sentences):

$$\mathcal{Y} = \{1, 2, ..., N\}$$

- 2^N possible subsets
- Probability measure $\mathcal{P}$ over subsets $Y \subseteq \mathcal{Y}$

Determinantal point process (DPP)

$$L = \begin{pmatrix} L_{11} & L_{12} & L_{13} & L_{14} \\ L_{21} & L_{22} & L_{23} & L_{24} \\ L_{31} & L_{32} & L_{33} & L_{34} \\ L_{41} & L_{42} & L_{43} & L_{44} \end{pmatrix}$$

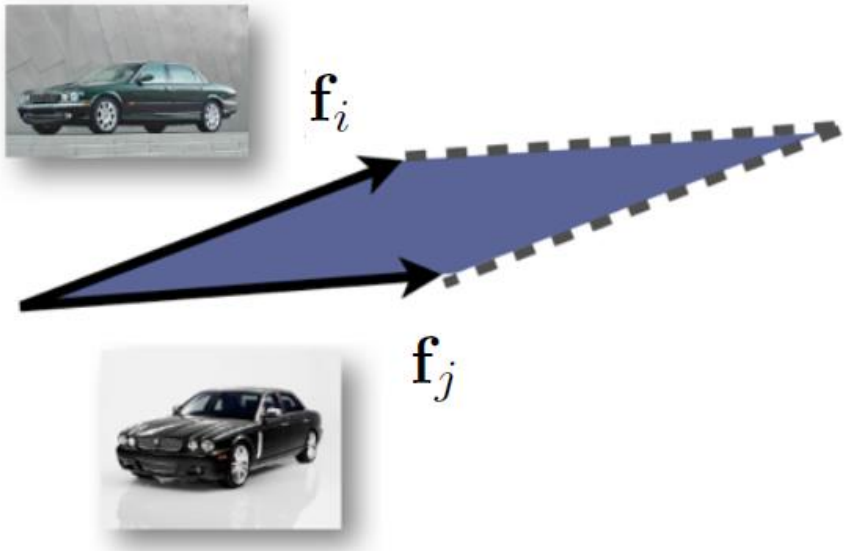
$$\mathcal{P}(Y) \propto \det(L_Y)$$

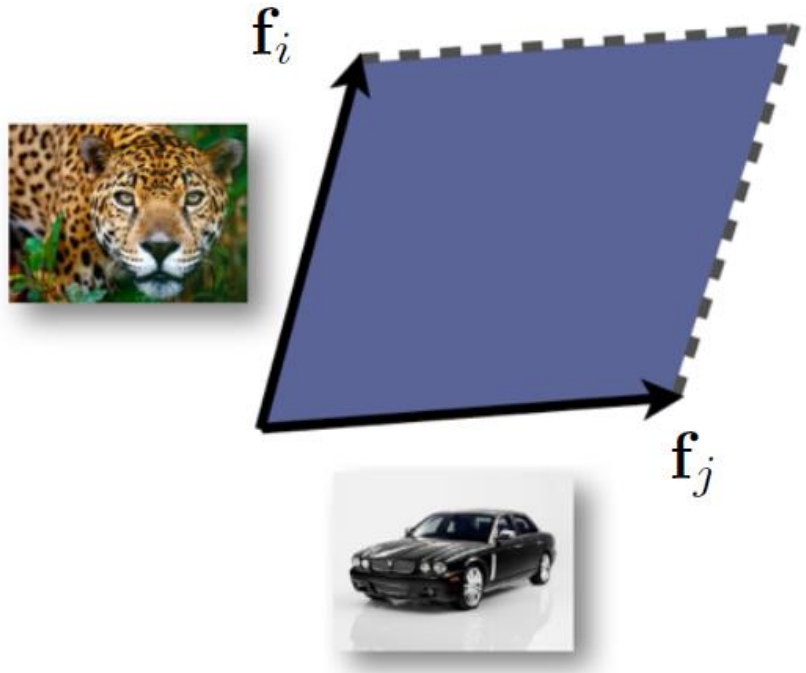
$$\mathcal{P}(\{2,4\}) \quad \begin{matrix} L_{11} & L_{12} & L_{13} & L_{14} \\ L_{21} & L_{22} & L_{23} & L_{24} \\ L_{31} & L_{32} & L_{33} & L_{34} \\ L_{41} & L_{42} & L_{43} & L_{44} \end{matrix}$$

$$\mathcal{P}(\{2,4\}) \propto \begin{vmatrix} L_{22} & L_{24} \\ L_{42} & L_{44} \end{vmatrix}$$

Why DPP models diversity?

If $S_{ij} = \langle \mathbf{f}_i, \mathbf{f}_j \rangle$





Determinant = the square of the area (2D), volume (3D), etc.

Sampling from DPP

Theorem 1. *Let $L = \sum_{n=1}^{N} \lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^\top$ be an orthonormal eigen-decomposition of a positive semidefinite matrix L, and let $\boldsymbol{e}_i$ be the ith standard basis N-vector (all zeros except for a 1 in the ith position). Then Algorithm 1 samples $\boldsymbol{Y} \sim \mathcal{P}_L$.*

Algorithm 1 Sampling from a DPP

Input: eigenvector/value pairs $\{(\boldsymbol{v}_n, \lambda_n)\}$

$J \leftarrow \emptyset$

for $n = 1, \ldots, N$ **do**

  $J \leftarrow J \cup \{n\}$ with prob. $\frac{\lambda_n}{\lambda_n + 1}$

end for

$V \leftarrow \{\boldsymbol{v}_n\}_{n \in J}$

$Y \leftarrow \emptyset$

while $|V| > 0$ **do**

  Select y_i from $\mathcal{Y}$ with $\Pr(y_i) = \frac{1}{|V|} \sum_{\boldsymbol{v} \in V} (\boldsymbol{v}^\top \boldsymbol{e}_i)^2$

  $Y \leftarrow Y \cup y_i$

  $V \leftarrow V_\perp$, an orthonormal basis for the subspace of V orthogonal to $\boldsymbol{e}_i$

end while

Output: Y

Sampling from k-DPPs

Algorithm 2 Sampling from a k-DPP

Input: eigenvector/value pairs $\{(\boldsymbol{v}_n, \lambda_n)\}$, size k

$J \leftarrow \emptyset$

for $n = N, \ldots, 1$ **do**

  if $u \sim U[0,1] < \lambda_n \frac{e_{k-1}^{n-1}}{e_k^n}$ **then**

    $J \leftarrow J \cup \{n\}$

    $k \leftarrow k - 1$

    if $k = 0$ **then**

      break

    end if

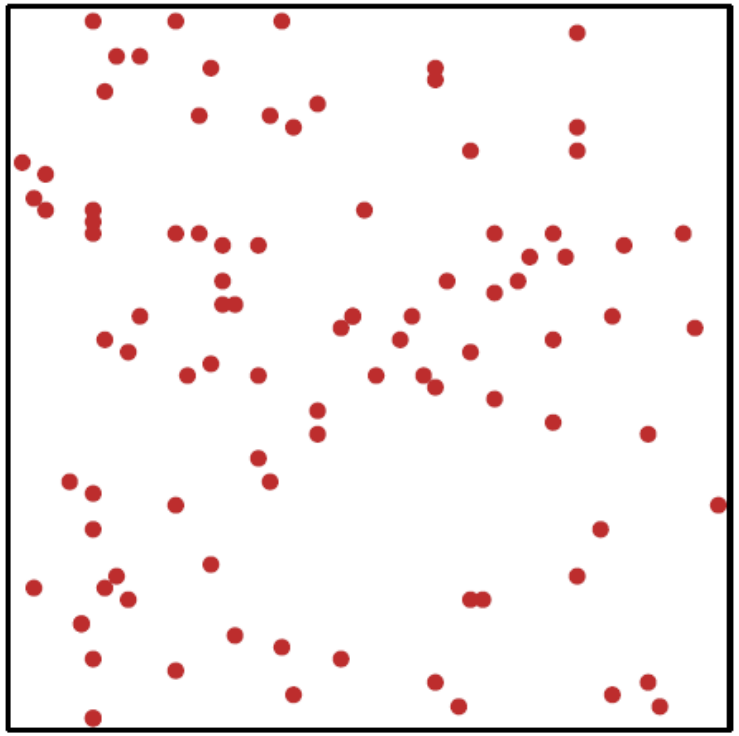
  end if

end for

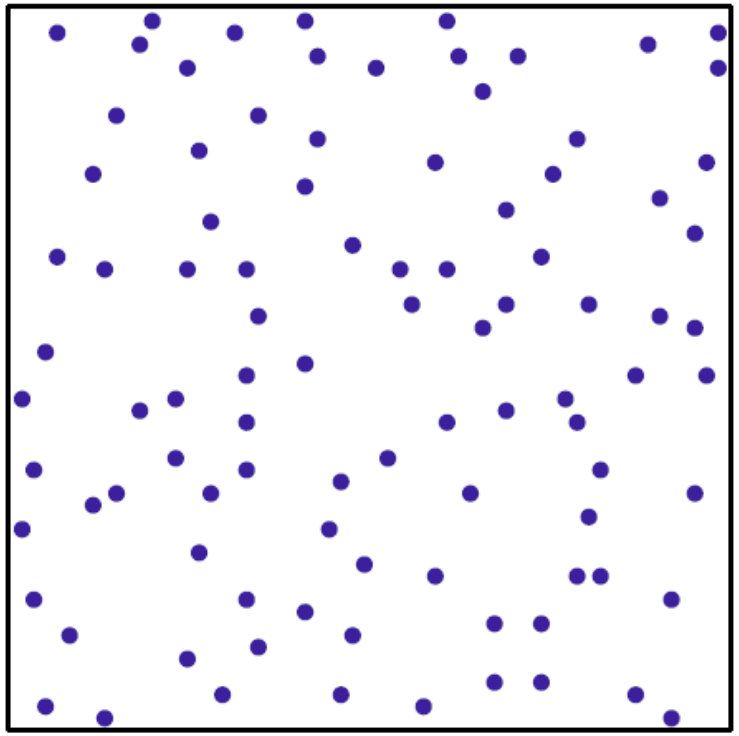
Proceed with the second loop of Algorithm 1

Output: Y

Point process samples



Independent



DPP

Image Retrieval

"jaguar"

By relevance

By relevance & diversity

Mini-batch Diversification

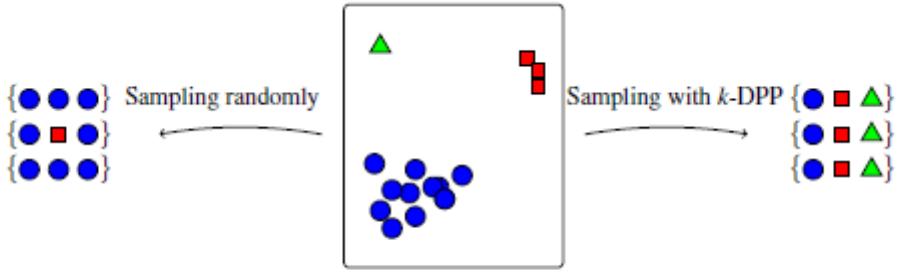


Figure 1: Sampling mini-batches using the k-DPP. For an imbalanced dataset, our method results in diversified mini-batches.

Mini-batch Diversification

Expected risk $J(\theta) = \mathop{\mathbb{E}}_{x \sim p_{\text{data}}} [\ell(x; \theta)]$

Empirical risk $\hat{J}(\theta) = \mathop{\mathbb{E}}_{x \sim p_{\text{emp}}} [\ell(x; \theta)] = \frac{1}{N} \sum_{i=1}^{N} \ell(x_i, \theta).$

Diversified risk $J^*(\theta) = \frac{1}{k} \mathop{\mathbb{E}}_{\vec{x} \sim \text{k-DPP}} [\ell(\vec{x}; \theta)],$

Algorithm

Algorithm 1 DM-SGD

Input: Data X, mini-batch size k, eigendecomposition $\{(v_n, \lambda_n)\}_{n=1}^{N}$ of similarity matrix K.

for $t = 0$ *to MaxIter* **do**

  Sample a mini-batch using the k-DPP

  Sample k eigenvectors V using eigenvalues;

  Sample mini-batch $\vec{x}$ of size k using V. (See supplement.)

  Update parameters

  $\theta_{t+1} = \theta_t + \rho_t g^*(\theta_t; \vec{x})$ (g* is the gradient estimate)

end

Algorithm 2 DM-SVI

We adopt the notation from [13].

for $t = 0$ *to MaxIter* **do**

  Sample a mini-batch using the k-DPP;

  Update variational parameters;

  for $j = 0$ *to Mini-batch Size* **do**

    Update local variational parameters (e.g. ϕ and λ for LDA) for mini-batch.

  end

  Compute the intermediate global parameters as if the mini-batch is replicated $\frac{D}{S}$ times.

  (e.g. $\tilde{\lambda}_{kw} = \eta + \frac{D}{S}\sum_{s=1}^{S} n_{tw}\phi_{twk}$ for LDA)

  Update the current estimate of the global variational parameters with $\rho_t = (\tau_0 + t)^{-k}$.

  $\lambda = (1 - \rho_t)\lambda + \rho_t \tilde{\lambda}$

end

Topic Learning With LDA

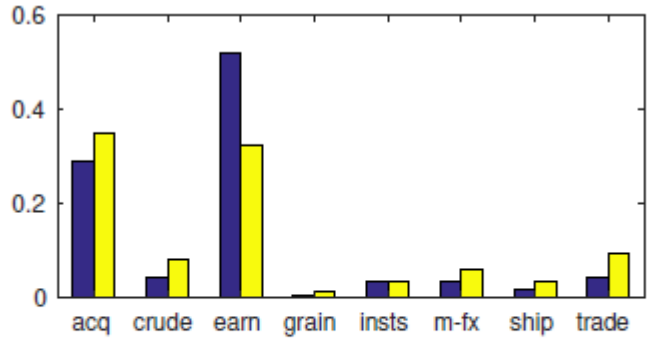
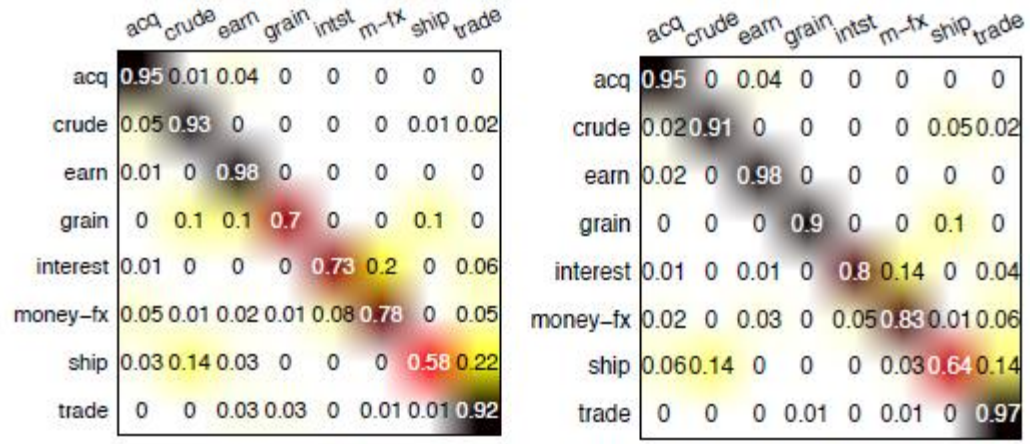


Figure 4: The frequency of class labels of the training dataset (in blue) and of the balanced dataset (in yellow). While explicit class label information is withheld, the algorithm partially balances class contributions.



(a) CM w.SVI (b) CM w.**DM-SVI**.

Figure 5: Confusion matrix for text classification based on LDA features obtained from SVI (a) and the proposed DM-SVI (b). DM-SVI features lead to better accuracies.

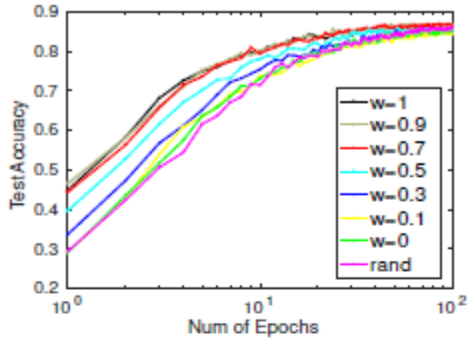
Multi-Class Logistic Regression

- Dataset: Oxford 102 flower
- Balanced training set and imbalanced test set
- use the original testing set for training and use the original training set for testing

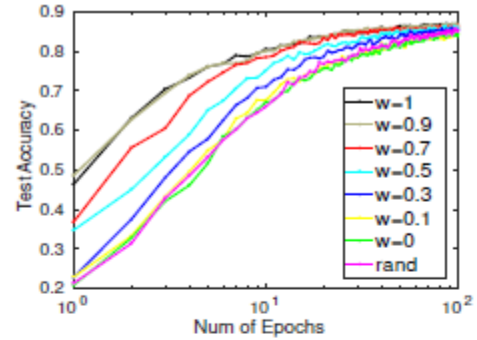
$$L = FF^{\top}$$

$$F = [(1-w)X_{fc1} \quad wH], \quad 0 \le w \le 1$$

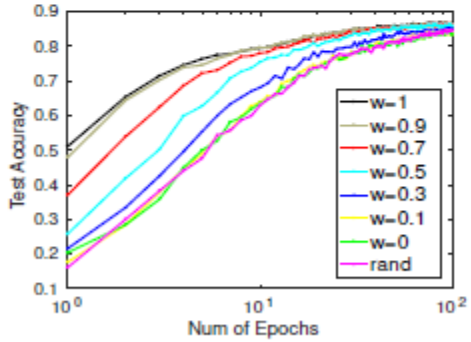
- When w is large, the algorithm focuses more on the class labels.
- When w is small, balancing is performed mostly based on the features.



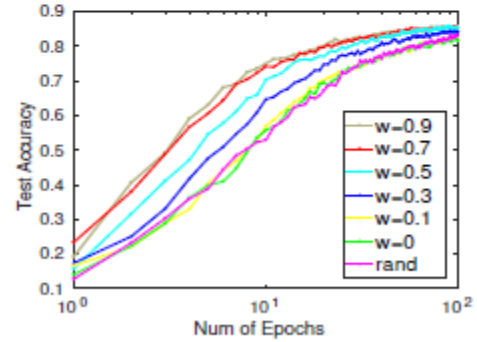
(a) k=50, Top3: 0.9, 0.7, 1;
Best: 86.7% Baseline:84.7%



(b) k=80, Top3: 0.9, 0.7, 1;
Best: 86.7% Baseline:81.8%

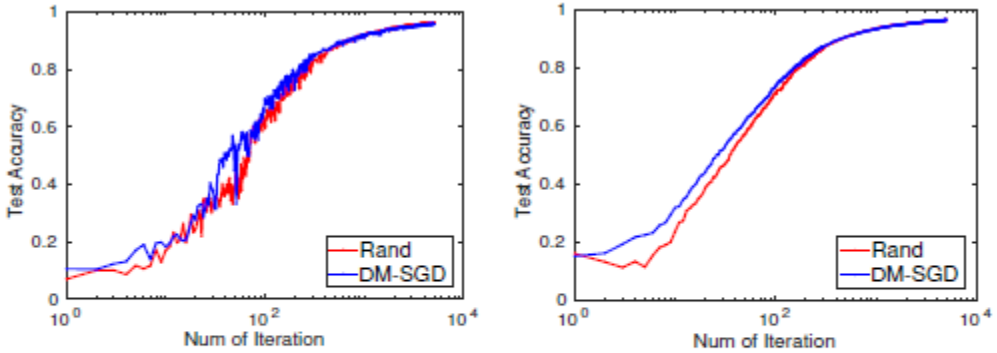


(c) k=102, Top3: 1, 0.9, 0.7;
Best: 86.5% Baseline:84.5%



(d) k=150, Top3: 0.7, 0.5, 0.9;
Best: 85.5% Baseline:83.1%

CNN Classification on MNIST



(a) k=10 (b) k=200

Figure 9: Same quantities shown as in Fig. 8, but for the MNIST data set, which is more balanced.

Diversifying Convex Transductive Experimental Design for Active Learning

Lei Shi[1,2] and **Yi-Dong Shen**[1]*
[1]State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences, Beijing 100190, China
{shilei,ydshen}@ios.ac.cn

IJCAI, 2016

Introduction

- A representative active learning method
- It uses a data reconstruction framework to select informative samples for labeling, where the informativeness of each sample is measured by its capacity to reconstruct the target data set.
- CTED: Assign each sample a score, which indicates the sample's capacity to reconstruct the target data set
- Similar samples may get similar ranking scores
- Diversifying CTED: Impose a diversity regularizer

Convex Transductive Experimental Design

representativeness of i-th example

CTED

$$\min_{\mathbf{A},\mathbf{b}} \quad ||\mathbf{X} - \mathbf{XA}||_F^2 + \sum_{i=1}^{n} \frac{\sum_{j=1}^{n} a_{ij}^2}{b_i} + \gamma||\mathbf{b}||_1$$
$$s.t. \quad b_i \geq 0, i = 1, 2, ..., n$$

A: reconstruction coefficients
b: sample selection vector

similar samples may get similar ranking scores

Diversifying CTED

$$\min_{\mathbf{A},\mathbf{b}} \quad ||\mathbf{X}-\mathbf{XA}||_F^2 + \sum_{i=1}^{n} \frac{\sum_{j=1}^{n} a_{ij}^2}{b_i} + \gamma||\mathbf{b}||_1 + \boxed{\alpha \mathbf{b}^T \mathbf{S} \mathbf{b}}$$
$$s.t. \quad b_i \geq 0, i = 1, 2, ..., n \qquad (3)$$

If s_{ij} is large, b_i and b_j cannot be large at the same time.

$$\min_{\mathbf{A},\mathbf{b}} \quad h(\mathbf{A},\mathbf{b}) = ||\mathbf{X}-\mathbf{XA}||_F^2 + \sum_{i=1}^{n} \frac{\sum_{j=1}^{n} a_{ij}^2}{b_i} + \gamma||\mathbf{b}||_1$$
$$+ \alpha \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\mathbf{a}^i(\mathbf{a}^j)^T}{||\mathbf{a}^i||_2||\mathbf{a}^j||_2} b_i b_j$$
$$s.t. \quad b_i \geq 0, i = 1, 2, ..., n \qquad (4)$$

$\mathbf{a}^i$ encodes the reconstruction coefficients of all the samples based on the i-th one.

$$s_{ij} = \frac{\mathbf{a}^i(\mathbf{a}^j)^T}{||\mathbf{a}^i||_2||\mathbf{a}^j||_2}$$

Optimization

Update $\widehat{\mathbf{A}}$

$$\min_{\widehat{\mathbf{A}}} \quad f(\widehat{\mathbf{A}}) = ||\mathbf{X} - \mathbf{X}\mathrm{diag}(\mathbf{s})\widehat{\mathbf{A}}||_F^2 + \alpha \mathrm{Tr}(\widehat{\mathbf{A}}^T \mathbf{b}\mathbf{b}^T \widehat{\mathbf{A}})$$
$$s.t. \quad ||\widehat{\mathbf{a}}^i||_2 = 1, i = 1, 2, ..., n$$

Update b

$$\min_{\mathbf{b}} \quad \sum_{i=1}^{n} \frac{s_i^2}{b_i} + \gamma ||\mathbf{b}||_1 + \alpha \mathbf{b}^T \widehat{\mathbf{A}} \widehat{\mathbf{A}}^T \mathbf{b}$$
$$s.t. \quad b_i \geq 0, i = 1, 2, ..., n$$

Update s

$$\min_{\mathbf{s}} \quad ||\mathbf{X} - \mathbf{X}\mathrm{diag}(\mathbf{s})\widehat{\mathbf{A}}||_F^2 + \sum_{i=1}^{n} \frac{s_i^2}{b_i}$$
$$s.t. \quad s_i \geq 0, i = 1, 2, ..., n$$

Experiment

Table 2: Results on USPS

	SVM			KNN		
m	C	D_f	D	C	D_f	D
10	0.578	0.747	**0.785**	0.595	0.694	**0.754**
20	0.661	**0.825**	**0.836**	0.658	0.776	**0.806**
30	0.705	**0.849**	**0.859**	0.716	**0.811**	**0.833**
40	0.760	0.864	**0.873**	0.765	**0.832**	**0.848**
50	0.824	**0.874**	**0.884**	0.815	0.849	**0.862**
60	0.854	0.878	**0.890**	0.834	**0.861**	**0.873**
70	0.858	0.882	**0.894**	0.846	0.868	**0.884**
80	0.864	0.884	**0.898**	0.857	0.872	**0.890**
90	0.870	0.887	**0.901**	0.858	0.875	**0.894**
100	0.872	0.890	**0.902**	0.867	0.879	**0.902**

Table 3: Results on MNIST

	SVM			KNN		
m	C	D_f	D	C	D_f	D
10	0.269	**0.427**	**0.426**	0.277	0.400	**0.431**
20	0.357	**0.537**	**0.549**	0.363	**0.507**	**0.527**
30	0.476	**0.600**	**0.609**	0.466	0.569	**0.591**
40	0.562	**0.633**	**0.643**	0.553	**0.603**	**0.622**
50	0.626	0.655	**0.683**	0.605	0.627	**0.661**
60	0.661	0.676	**0.711**	0.636	0.644	**0.682**
70	0.699	0.691	**0.727**	0.665	0.659	**0.701**
80	0.720	0.696	**0.740**	0.687	0.671	**0.715**
90	**0.735**	0.704	**0.750**	0.700	0.681	**0.724**
100	**0.752**	0.710	**0.760**	0.712	0.689	**0.736**

Table 4: Results on NewsGroup

	SVM			KNN		
m	C	D_f	D	C	D_f	D
10	0.581	0.611	**0.626**	0.575	**0.596**	**0.614**
20	0.690	0.684	**0.720**	0.663	0.662	**0.694**
30	0.732	0.742	**0.770**	0.690	0.700	**0.717**
40	0.760	0.776	**0.802**	0.715	0.719	**0.736**
50	0.788	0.793	**0.820**	0.730	0.733	**0.749**
60	0.815	0.808	**0.832**	0.742	0.741	**0.760**
70	**0.832**	**0.830**	**0.844**	0.752	0.754	**0.767**
80	0.840	0.838	**0.856**	0.758	0.762	**0.774**
90	0.846	0.847	**0.861**	0.763	0.766	**0.780**
100	**0.853**	**0.853**	**0.863**	**0.772**	0.769	**0.780**

Table 5: Results on WEBKB

	SVM			KNN		
m	C	D_f	D	C	D_f	D
10	0.595	**0.626**	**0.634**	0.555	**0.609**	**0.609**
20	0.659	**0.685**	**0.693**	0.601	**0.636**	**0.641**
30	0.703	**0.716**	**0.725**	0.632	**0.650**	**0.651**
40	0.716	0.740	**0.754**	0.641	**0.661**	**0.659**
50	0.722	**0.761**	**0.774**	0.647	**0.666**	**0.668**
60	0.732	**0.776**	**0.784**	0.657	**0.671**	**0.675**
70	0.743	0.782	**0.799**	0.659	**0.673**	**0.679**
80	0.753	0.797	**0.808**	0.665	**0.679**	**0.683**
90	0.758	0.806	**0.821**	0.670	**0.684**	**0.688**
100	0.765	0.815	**0.828**	0.674	**0.686**	**0.691**

Table 6: Results on ORL

	SVM			KNN		
m	C	D_f	D	C	D_f	D
10	0.167	0.185	**0.221**	0.163	0.175	**0.211**
20	0.306	0.313	**0.346**	0.274	0.285	**0.327**
30	0.396	**0.418**	**0.439**	0.360	0.377	**0.411**
40	0.467	0.475	**0.519**	0.430	0.436	**0.479**
50	0.524	0.537	**0.589**	0.481	0.495	**0.544**
60	0.562	0.570	**0.636**	0.517	0.525	**0.586**
70	0.601	0.614	**0.698**	0.555	0.567	**0.628**
80	0.622	0.642	**0.728**	0.582	0.604	**0.668**
90	0.666	0.680	**0.766**	0.616	0.628	**0.705**
100	0.700	0.716	**0.786**	0.641	0.655	**0.728**

Experiment

To evaluate the effectiveness of the proposed Diversified CTED (DCTED for short), we compare DCTED with the following closely related methods, including distribution matching via Maximum Mean Discrepancy (MMD) [Chattopadhyay *et al.*, 2012], Convex Transductive Experimental Design (CTED) [Yu *et al.*, 2008], Active Learning via Neighbourhood Reconstruction (ALNR) [Hu *et al.*, 2013], and Accelerated Robust Subset Selection (ARSS) [Zhu and Fan, 2015]. We also compare DCTED with the method corresponding to Eq. (3). This method leverages a pre-defined and fixed similarity matrix. We denote this method as DCTED_f. For

Table 7: Results Based on SVM

Methods	USPS	MNIST	NewsG	WEBKB	ORL
Random	0.718	0.565	0.695	0.701	0.497
CTED	0.784	0.586	0.774	0.715	0.501
ALNR	0.813	0.492	0.763	0.712	0.494
ARSS	0.847	0.614	0.765	0.715	0.517
MMD	0.841	0.640	0.755	0.719	0.494
DCTED_f	0.858	0.633	0.778	0.750	0.515
DCTED	**0.872**	**0.660**	**0.799**	**0.762**	**0.573**

Table 8: Results Based on KNN Classifier

Methods	USPS	MNIST	NewsG	WEBKB	ORL
Random	0.710	0.590	0.638	0.613	0.453
CTED	0.781	0.566	0.716	0.640	0.462
ALNR	0.819	0.475	0.717	0.637	0.456
ARSS	**0.849**	0.590	0.705	0.639	0.475
MMD	0.838	0.617	0.696	0.625	0.450
DCTED_f	0.832	0.605	0.720	**0.662**	0.474
DCTED	**0.855**	**0.639**	**0.737**	**0.664**	**0.528**