行列上のガウス分布を用いた Confidence Weighted Linear Classifierの多クラス化

大岩秀和, 松島慎, 中川裕志 (東京大学)

分類問題

- あるデータxが与えられたとき、データxがどのカテゴリyに属するかを判定する問題

$$\mathbf{x} \in \Re^D \xrightarrow{\text{yを予測}} h(\mathbf{x}) \in \{1,2,\ldots,K\} = y\,?$$

$\mathbf{x}$: データを表す特徴ベクトル(D次元)

y : データの所属するクラス

$h(\cdot)$: 識別関数

- 特に、K≧3の分類問題を多クラス分類問題と呼ぶ
- 精度の高い識別関数h(・)を出来るだけ高速に求めたい

本研究では、識別関数を線形の場合に限定

$$h(\mathbf{x}) = \arg\max_{y \in \{1,\ldots K\}} \left(\mathbf{w} \cdot f(\mathbf{x}, y)\right)$$

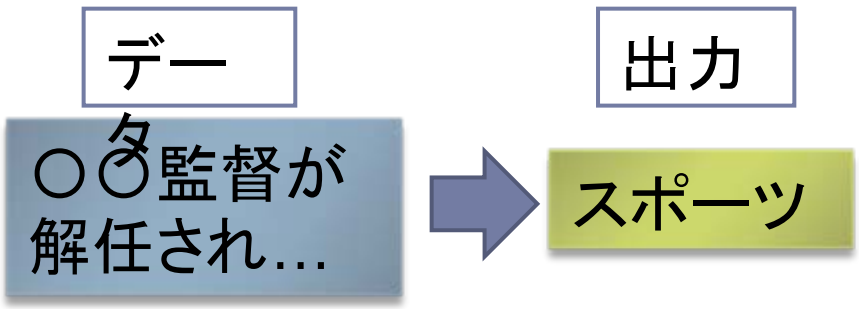
$$f(\mathbf{x}, y) = (\mathbf{x}^T \delta_{y=1}, \ldots, \mathbf{x}^T \delta_{y=K})^T$$

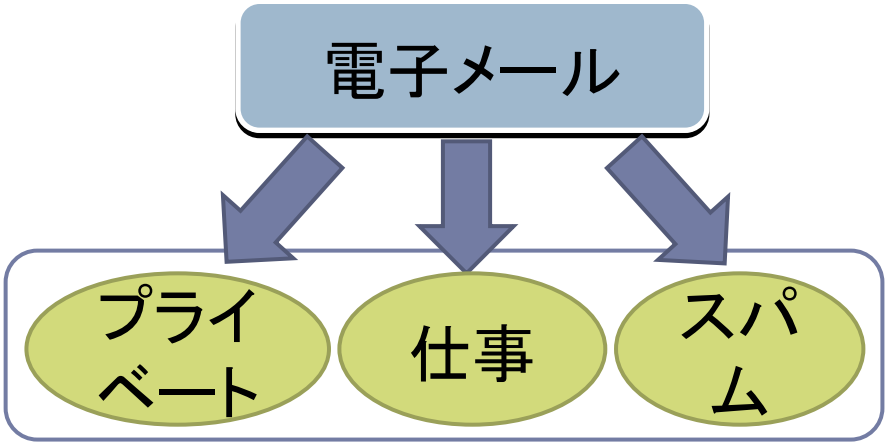
$\mathbf{w}$: 重みベクトル

各クラスのスコア : $s_y = \mathbf{w} \cdot f(\mathbf{x}, y)$ ⇨ 識別関数 : $h(\mathbf{x}) = \arg\max_y (s_y)$

分類問題の応用

- ニュースのカテゴリ分類
  - 文書　⇒　スポーツ？経済？政治？
- テキストの言語判定
  - 文書　⇒　英語？フランス語？ドイツ語？
- メールの分類
  - メール⇒　仕事？プライベート？スパム？





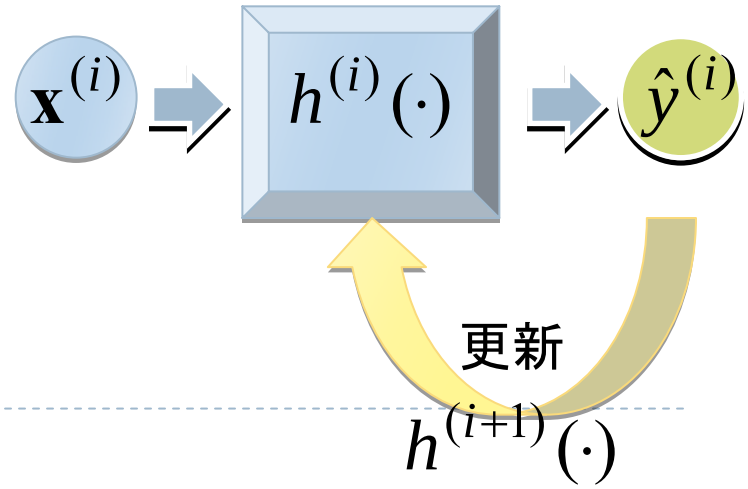
オンライン学習

- データが1つ来るたびに、識別関数h(・)を逐次的に更新
  - データ $\mathbf{x}^{(i)}$ を受け取る
  - 識別関数 $h^{(i)}(\cdot)$ を用いて、データ $\mathbf{x}^{(i)}$ の所属するクラス $\hat{y}^{(i)}$ を予測

$$\hat{y}^{(i)} = h^{(i)}(\mathbf{x}^{(i)})$$

  - 正解クラス $y^{(i)}$ と $\hat{y}^{(i)}$ を比較し、クラスの予測が上手く出来ていない場合は、識別関数を更新 $h^{(i)}(\cdot) \mapsto h^{(i+1)}(\cdot)$
  - 上記の操作を繰り返す

- オンライン学習の特徴
  - 収束が高速
  - 大量のメモリを必要としない

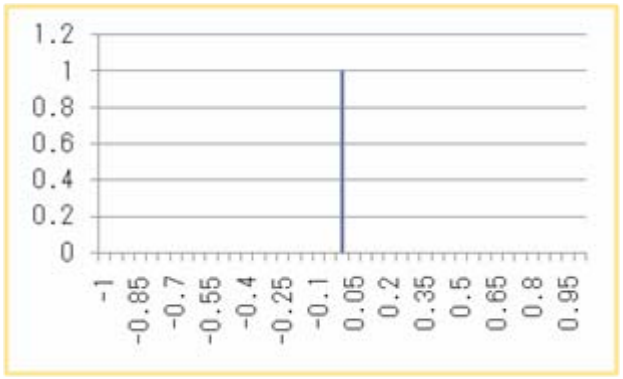


Multi-Class Confidence Weighted Algorithms(MCCW)

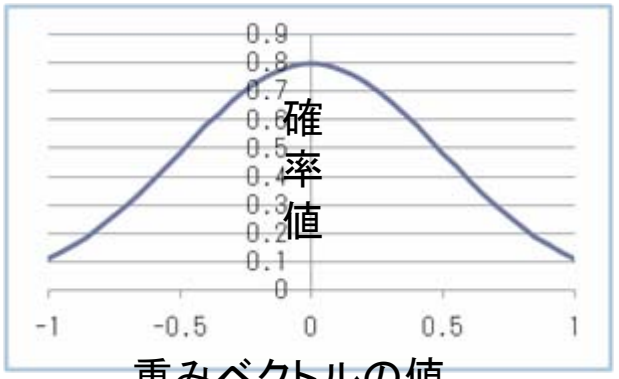
[Crammer et al. 2009]

- 多クラス分類問題に対するオンライン学習手法
  - 特に自然言語系のタスクに対して、高い識別性能を示す
- 重みベクトル上にガウス分布を導入
  - 出現頻度の低い特徴には大きな更新幅を与えるアルゴリズムを実現
  - “Confidence-weighted linear classification” [Dredze et al. 2008]の多クラス分類問題への拡張

既存手法



MCCW



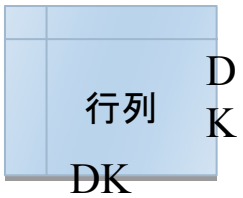
$$h(\mathbf{x}) = \arg\max_y (s_y)$$

$$h(\mathbf{x}) = \arg\max_y E[s_y]$$

MCCWの定式化

- 重みベクトルに直接ガウス分布を導入

$$\mathbf{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_{DK} \end{pmatrix} \mapsto \mathbf{w} \sim N(\boldsymbol{\mu}, \Sigma)$$



パラメータ数
が多すぎる…

ガウス分布を導入

$\boldsymbol{\mu}$: 重みベクトルの平均(DK次元)

Σ : 重みベクトルの共分散行列($DK \times DK$次元)

- 最適化問題

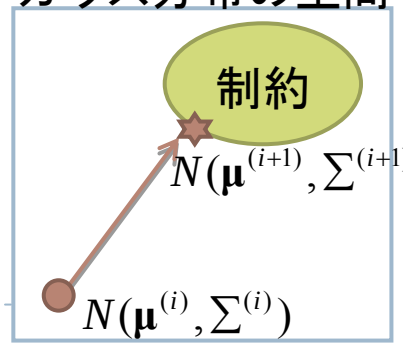
現在の多変量ガウス分布に
最も近いガウス分布を選択

$$\min D_{KL}\left(N(\boldsymbol{\mu}, \Sigma) \| N(\boldsymbol{\mu}^{(i)}, \Sigma^{(i)})\right)$$

$$s.t. \quad \Pr_{\mathbf{w} \sim N(\boldsymbol{\mu}, \Sigma)}\left[s_{y^{(i)}} \geq s_k\right] \geq \eta \quad \left(\forall k \neq y^{(i)}\right)$$

正解クラスyが各クラスkに対して、
η 以上の確率で正しく識別する制約

ガウス分布の空間



MCCWの性質

- Σの次元数が大きいため、非対角項を0と置く

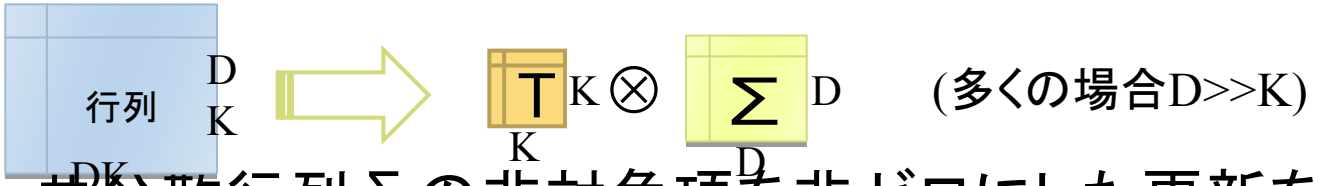
- 制約式を緩和する(上位数ラベルのみ)
  - Single Constraint (k=1)
    - 正解クラスと最も平均スコアの高いクラスとの2値分類問題
  - Sequential Constraint (k=∞)
    - 正解クラスと平均スコア上位の数クラスそれぞれとの2値分類問題をスコアの高い順に解く
    - 全クラスについて制約を課すと過学習が発生する事が指摘されている
      - 正解クラスの優遇しすぎ
      - 不正解クラスの叩き過ぎ
  - Parallelな手法も存在

提案手法

提案１ – 重み行列上にガウス分布を導入

更新に必要なパラメータ数削減

特徴間の共分散行列とクラス間の共分散行列を分離



共分散行列Σの非対角項を非ゼロにした更新を可能に

提案２ - サポートクラスの導入

サポートクラス -- 制約が有効に働くクラス

効率的なクラス選択によって、近似解法を必要としない

全クラスの制約を同時に考慮した厳密な更新を実現

過学習の問題を緩和

提案1:重み行列上へのガウス分布の導入

- 重みベクトルに直接ガウス分布を導入せず、重みベクトルを行列の形に直し、重み行列上にガウス分布を導入

重みベクトル

$$\mathbf{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_{DK} \end{pmatrix} \mapsto \mathbf{W} = \begin{pmatrix} w_{11} & \cdots & w_{N1} \\ \vdots & \ddots & \vdots \\ w_{1D} & \cdots & w_{ND} \end{pmatrix}$$

重み行列

ガウス分布を導入

$$\mapsto \mathbf{W} \sim \mathrm{N}_{D\times K}\left((\mu_1, \mu_2, \ldots, \mu_K), T, \Sigma\right)$$

Σ D D

μ_k :クラスkに対応する重みベクトルの平均(D次元)

T :クラス間の関係を表す列共分散行列($K \times K$次元)

Σ :特徴間の関係を表す行共分散行列($D \times D$次元)

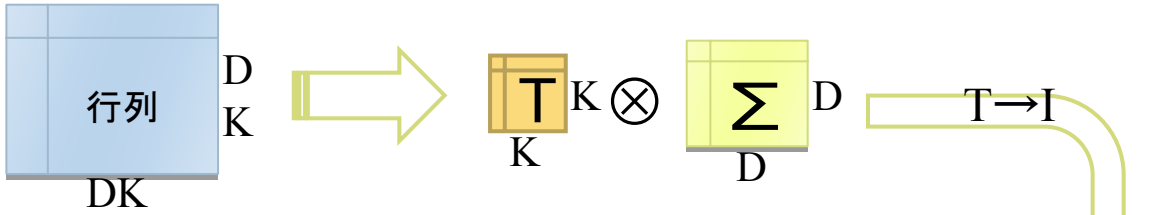
重み行列を重みベクトルに戻す

重み行列

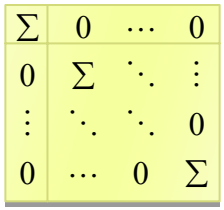
$$\mathbf{W} = \begin{pmatrix} w_{11} & \cdots & w_{N1} \\ \vdots & \ddots & \vdots \\ w_{1D} & \cdots & w_{ND} \end{pmatrix} \mapsto \mathbf{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_{DK} \end{pmatrix}$$

重みベクトル

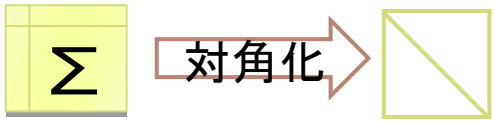
$$\mathbf{W} \sim \mathrm{N}_{D\times K}\left((\mu_1,\ldots,\mu_K),T,\Sigma\right) \mapsto \mathbf{w} \sim \mathrm{N}\left((\mu_1,\ldots,\mu_K),T\otimes\Sigma\right)$$



- 上記の操作によって、パラメータ数が減少
  - クラス間の関係Tは事前に予測可能
  - 本研究ではTを単位行列Iに近似→D×D次元
  - Σの非対角項を0とおく実験も行う(SCCWD)



SCCW
(Support Class Confidence Weighted)



SCCWD

更新式の導出

現在の多変量ガウス分布に
最も近いガウス分布を選択

- 最適化問題

$$(\boldsymbol{\mu}^{(i+1)}, \Sigma^{(i+1)}) = \arg\min_{\boldsymbol{\mu},\Sigma} D_{KL}\left(N(\boldsymbol{\mu}, \Sigma \otimes I) \middle\| N(\boldsymbol{\mu}^{(i)}, \Sigma^{(i)} \otimes I)\right)$$

$$s.t. \quad \Pr_{\mathbf{w}\sim N(\boldsymbol{\mu},\Sigma\otimes I)}\left[s_{y^{(i)}} \geq s_k\right] \geq \eta \quad \left(\forall k \neq y^{(i)}\right)$$

正解クラスyが各クラスkに対して、
η 以上の確率で正しく識別する制約

- パラメータ更新式

$$\boldsymbol{\mu}_k^{(i+1)} = \boldsymbol{\mu}_k^{(i)} - \alpha_k^{(i)} \Sigma^{(i)} \mathbf{x}^{(i)} \quad (\forall k \neq y^{(i)})$$

$$\boldsymbol{\mu}_k^{(i+1)} = \boldsymbol{\mu}_k^{(i)} + \sum_{k\neq y^{(i)}} \alpha_k^{(i)} \Sigma^{(i)} \mathbf{x}^{(i)} \quad (k = y^{(i)})$$

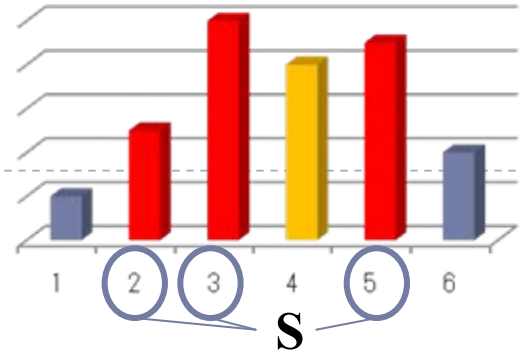
ϕ : 標準正規分布の累積密度関数

$$\Phi = \phi^{-1}(\eta)$$

$$\Sigma^{(i+1)} = \Sigma^{(i)} - \Sigma^{(i)} \mathbf{x}^{(i)} \left(\frac{\sqrt{2}\Phi \sum_{k\neq y^{(i)}} \alpha_k^{(i)}}{K\sqrt{(\mathbf{x}^{(i)})^T \Sigma^{(i+1)} \mathbf{x}^{(i)}} + \sqrt{2}\Phi \sum_{k\neq y^{(i)}} \alpha_k^{(i)} \left((\mathbf{x}^{(i)})^T \Sigma^{(i)} \mathbf{x}^{(i)}\right)} \right) (\mathbf{x}^{(i)})^T \Sigma^{(i)}$$

Lagrange乗数 α_kが求まれば、更新式が閉じた形で記述可能

提案2:サポートクラスの導入



- サポートクラス
  - 制約式が有効に働くクラス
  - KKT条件にて、$\Pr_{\mathbf{w}}[s_{y^{(i)}} \geq s_k] = \eta$が成立するクラスk

- サポートクラスを導入すると、先の最適化問題の α_kを、近似解法を用いずに、効率的に求めることが出来る

$$\sum_{s \neq y^{(i)}} \alpha_s^{(i)} = \frac{-KB\sum_{k\in S} l_k^{(i)} + |S|K\Phi\sqrt{(\Phi^2(v^{(i)})^2)\left(\sum_{k\in S} l_k^{(i)}\right)^2 + 2v^{(i)}\left(B^2 - |S|^2\Phi^4(v^{(i)})^2\right)}}{\left(B^2 - |S|^2\Phi^4(v^{(i)})^2\right)}$$

$$\alpha_k^{(i)} = \frac{-l_k^{(i)} + \Phi\sqrt{2u^{(i)}}}{v^{(i)}} - \sum_{s \neq y^{(i)}} \alpha_s^{(i)}$$

$|S|$:サポートクラスのクラス数

$$l_k = (\boldsymbol{\mu}_y - \boldsymbol{\mu}_k)\cdot \mathbf{x}$$

$$A = \sum_{s \neq y^{(i)}} \alpha_s^{(i)}$$

$$B = (|S|+1)Kv^{(i)} + |S|\Phi^2 v^{(i)}$$

実験

- 比較対象
  - Multi-Class Confidence Weighted(CW)
    - Single Constraint (k=1)
    - Sequantial Constraint (k=∞)
  - Passive-Aggressive(PA, PA-I, PA-II)
- 10-fold Cross-Validation(3 iteration)
- データセット

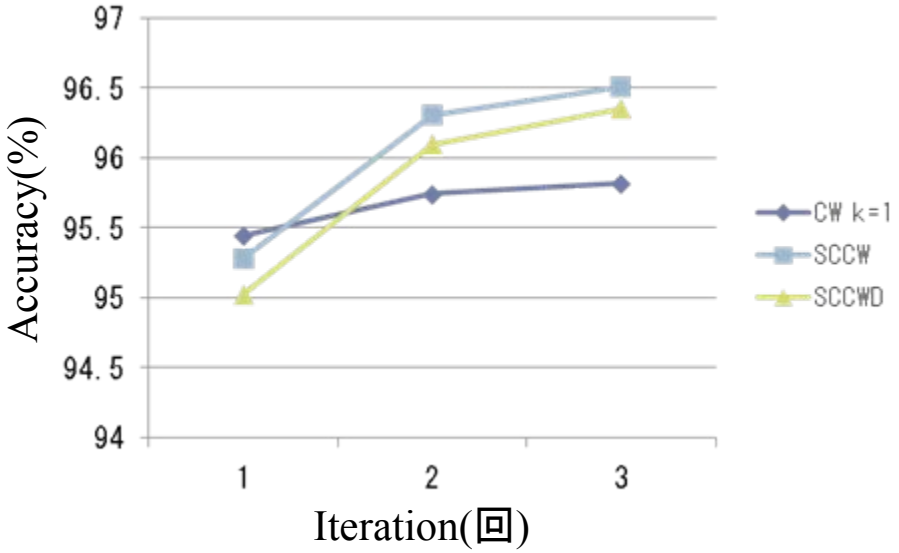
	概要	特徴次元数	クラス数
news20	ニュース記事のカテゴリ分類	60,345	20
~-7-1 **~-8-1**	news20のサブセット	9,605~16,282	7~8
reut20	ニュース記事のカテゴリ分類	34,488	20
USPS	手書き数字認識	256	10

実験結果

	PA	PA-I	PA-II	CW k=1	CW k=ALL	SCCW	SCCWD
ol-7-1	87.2	88.9	88.7	**94.1**	91.2	93.1	93.0
ol-8-1	87.9	89.7	89.7	**94.8**	92.0	94.0	94.0
ob-7-1	88.9	91.1	90.8	**95.6**	91.8	94.9	94.9
ob-8-1	90.3	90.4	90.5	**94.8**	92.0	94.6	94.7
sb-7-1	92.2	92.8	92.7	**96.5**	95.4	96.1	96.1
sb-8-1	92.2	92.5	92.8	**97.0**	94.8	96.3	96.3
reut20	94.4	94.4	94.2	95.8	93.9	**96.5**	96.4
USPS	90.0	89.6	90.4	92.4	83.5	**93.7**	93.5
news20	75.8	78.0	78.0	**84.8**	76.3	83.3	83.3

Reut20,USPSに対して最高精度を実現
一方、news20とそのサブセットに対してはCW k=1の精度が高い

反復回数毎の精度評価

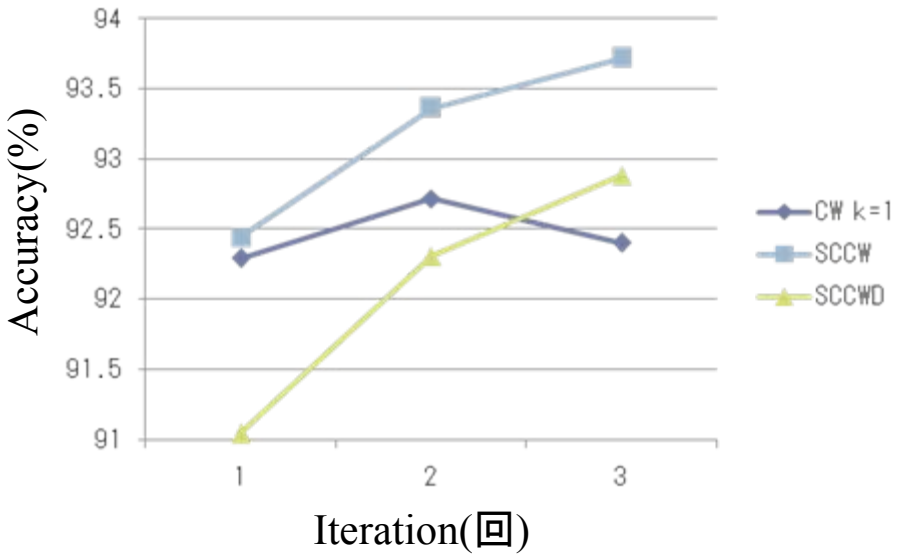


(a)reut20

Data Size:7,800

Feature dimension:34,488

Class:20



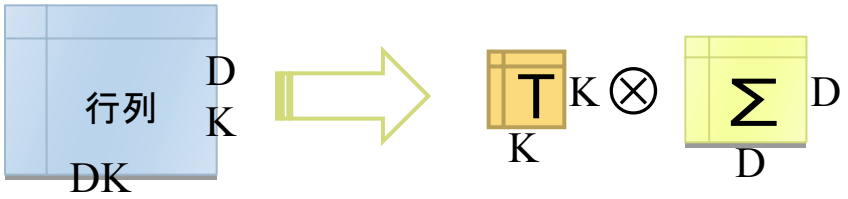
(b)USPS

Data Size:7,291

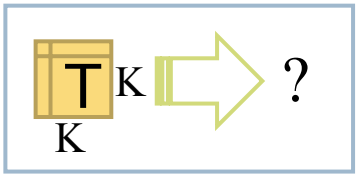
Feature dimension:256

Class:10

今後の課題



- クラス間の関係を表す行列Tのパラメータ更新
  - 各クラスの出現頻度に応じた、Tの対角成分の更新
  - クラス間の関係を非対角項を用いて更新
  - 上手くやれば、さらなる精度向上が望める



- 提案手法のMistake Bound Analysis
- 他のデータセット／既存手法に対する精度評価
  - よりクラス数の多いデータセット
  - MCCWのk=5 Sequential/Parallelに対する性能評価