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　　中川浩志（東京大学）

概要

本研究では、多クラス分類問題に対する新たな学習手法を提案する。

2009 年に Crammer らによって、Multi-class Confidence Weighted Algorithm(MCCW) [3] が提案された。MCCW は、自然言語の持つ性質を上手く捉えたアルゴリズムを実現しており、テキスト分類などの多クラス分類問題に対して、非常に高い識別精度を示している。しかし、この手法は、クラス数が増加すると、パラメータ数が非常に大きくなる問題が指摘されている。そのため、MCCW では、更新するパラメータを制限し、また最適化問題の制約式を減らすことで、計算量の問題を防いでいる。

本研究では、重み行列にガウス分布を導入し、最適化問題のパラメータ数を減少させることで、計算量の問題を緩和する手法を提案する。さらに、サポートクラス [4] と呼ばれる手法を導入し、厳密で効率的なパラメータ更新を行うアルゴリズムを提案する。

最後に、提案手法を実データに基づく多クラス分類問題に適用し、既存手法との精度比較を行う。

1　はじめに

テキストなどのデータが与えられたとき、そのデータがどのクラスに属するのかを判定する問題を分類問題と呼ぶ。特に、3 つ以上の候補の中から各データが所属するクラスを 1 つに決定する分類問題は、多クラス分類問題と呼ばれる。多クラス分類問題は、ニュース記事のカテゴリ分類やテキストの言語判定、手書き文字認識など、言語処理をはじめとする多くの分野で重要な研究課題となっている。そのため、多クラス分類問題に対する様々なアルゴリズムが研究されている。

また、オンライン学習と呼ばれる学習手法が近年注目を浴びている。全てのデータを一度に読み込んで学習を行う従来の機械学習の手法とは異なり、オンライン学習はデータが 1 つ与えられるたびに学習器を毎回更新していく手法である。オンライン学習手法は従来の学習手法に比べ、収束が速く、空間計算量も小さくなる性質を持つ。従って、大規模なデータセットに対しても高価な計算機を用いること無く高速に学習を行うことが出来るという利点があり、多クラス分類問題に対するオンライン学習アルゴリズムの研究が近年活発となっている。

ここで、テキストのカテゴリ分類などの自然言語を対象とした分類問題には、他の分類問題には無い性質を持つことを指摘しなければならない。自然言語を対象とする分類問題では、単語数が豊富であることから特徴ベクトルが非常に高次元となりやすく、かつ、スパースになる性質を持ちやすい。さらに、他の単語に比べ出現頻度の低い単語が所属するクラスを変化させる事も多い。

しかし、既存の多クラス分類問題に対するオンライン学習による学習手法には、各特徴の出現頻度を考慮してパラメータ更新を行うものは少なく、上記の自然言語の性質をとらえる事が難しかった。

2008 年に Dredze らによって、Confidence Weighted Linear Classifier(CW) [1, 2] と呼ばれる、2 値分類問題のためのオンライン学習アルゴリズムが提案された。CW は、重みベクトルにガウス分布を導入することで、低頻度の特徴にはパラメータの更新幅が大きくなるように設計されたアルゴリズムであり、自然言語の性質を取り込む事に成功した。

しかし、CW を多クラス分類問題に適用することは難しく、Crammer らが CW を多クラス分類問題へ拡張した Multi-class Confidence Weighted Algorithms(MCCW) [3] では、近似的な手法を用いることで、パラメータの更新式を導出している。さらに、共分散行列の次元数がクラス数の増加に従って非常に大きくなるため、非対角項を厳密に計算すると、計算量の問題が発生することが指摘されている。このように、自然言語の特徴を上手く捉えた多クラス分類問題に対するオンライン学習アルゴリズムには、まだ多くの研究余地が残されている。

そこで本研究では、CW を多クラス分類問題へ拡張する際に、重みベクトルを行列とみなすことで、パラメータ数を減少させる方法を提案する。さらに、サポートクラスと呼ばれる手法を導入することで、近似的な手法ではなく、厳密にパラメータ更新を行う手法を提案する。

この提案手法により、共分散行列の非対角項を計算する場合にも、クラス数に応じてパラメータ数が爆発しない更新を可能にし、さらに、サポートクラスを導入し、厳密なパラメータ更新を行うことで、MCCW で指摘されていた過学習の問題が緩和される事が期待される。

さらに、提案手法と既存手法について実データを用いて実験を行ない、精度を評価することで、提案手法と既存手法の性能を比較する。精度評価の結果、一部のデータセットに対して、提案手法が既存手法を上回る精度が確認されたことを示す。

本論文の構成は以下の通りである。はじめに第 2 章で、多クラス分類問題の問題設定について述べる。第 3 章では、先行研究として CW と MCCW を紹介する。第 4 章では提案手法について述べ、第 5 章では、実データを用いて既存手法と提案手法の性能比較を行う。最後に第 6 章で、本研究のまとめと今後の課題について述べる。

2 問題設定

多クラス識別問題とは、あるクラスに属するデータが D 次元の入力空間 $\mathbf{x} \in X \subset \Re^D$ の特徴ベクトルで表されるときに、$\mathbf{x}$ が所属するクラスを判定する関数 $h(\mathbf{x}) \in Y \subset \{1, 2, ..., K\}$ を推定する問題である。

また、本研究では、オンライン学習手法を用いる。オンライン学習では、以下の手順で識別関数の学習を行う。はじめに単一のデータ $\mathbf{x}^{(i)}$ が与えられ、$\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(i-1)}$ から構成した識別関数 $h^{(i)}$ を用いて、$\mathbf{x}^{(i)}$ が所属するクラス $\hat{y}^{(i)} = h^{(i)}(\mathbf{x}^{(i)})$ を推定する。次に、正解データ $y^{(i)}$ を受け取り、$\hat{y}^{(i)}$ と比較する。今回のデータが上手く分類されていない場合には、よりよい分類が可能になるように学習器を逐次的に更新する。

学習過程において、誤分類 $y^{(i)} \neq \hat{y}^{(i)}$ の回数を最小化することが学習器の目的となる。

さらに、本研究では、特徴ベクトルを $f(\mathbf{x}, y) = (\mathbf{x}^T\delta_{y=1}, \mathbf{x}^T\delta_{y=2}, ..., \mathbf{x}^T\delta_{y=K})^T \in \Re^{DK}$ と表記する。k が Y のあるクラスを表すとき、$\delta_{y=k}$ は y が k の時に 1、その他の値の場合は 0 となる関数である。

$$\delta_{y=k} = \begin{cases} 0 & (y \neq k) \\ 1 & (y = k) \end{cases}$$

本研究では、識別関数を線形関数に限定する。つまり、識別関数は重みベクトル $\mathbf{w}$ で特徴付けられた線形

モデルのみを対象とするため、以下の形で与えられるとする。

$$h_{\mathbf{w}}(\mathbf{x}) = \arg\min_{y \in Y}(\mathbf{w} \cdot f(\mathbf{x}, y)) \tag{1}$$

上式から分かるように、識別関数は入力ベクトル $\mathbf{x}$ が与えられたとき、$\mathbf{w} \cdot f(\mathbf{x}, y)$ が最大となるクラス y を $\mathbf{x}$ が所属するクラスとして予測する。

3 先行研究

本章では、識別問題に対するオンライン学習アルゴリズムの先行研究として Confidence Weighted Linear Classifier(CW) [1] を紹介する。さらに CW を多クラス分類問題に拡張した、Multi-class Confidence Weighted Algorithms(MCCW) [3] と呼ばれる手法を紹介する。

CW は、Dredze et al. によって提案された 2 値分類問題のためのオンライン学習アルゴリズムである。ここで 2 値分類問題とは、クラス集合が $Y = \{1, -1\}$ となる問題を指す。

CW は、入力ベクトルに対応する D 次元の重みベクトル $\mathbf{w}$ を、平均が $\boldsymbol{\mu} \in \Re^d$、共分散行列は $\Sigma \in \Re^{d \times d}$ で表される多変量ガウス分布 $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ とおいて、学習を行う。

CW では、入力ベクトル $\mathbf{x}^{(i)}$ が与えられたとき、$\mathbf{w}^{(i)} \cdot \mathbf{x}^{(i)}$ の符号に従って、入力ベクトルが所属するクラスを予測する。次に、入力ベクトルに対応する正解クラス $y^{(i)}$ が与えられ、$Pr_{\mathbf{w}^{(i)}}[y^{(i)}(\mathbf{w}^{(i)} \cdot \mathbf{x}^{(i)}) \geq 0] \geq \eta$ を求めることで、予測結果が頑健であるかどうかを調べる。ここで、η は、予測の頑健さの基準を $0.5 \leq \eta < 1$ の条件下で与えるための、閾値パラメータである。

もし与えられた入力ベクトル $\mathbf{x}^{(i)}$ が正しく分類されなかった場合、今回受け取ったデータを上手く識別できるようにパラメータを更新する条件のもと、重みベクトルが KL-divergence において最小の更新幅となるように、重みベクトル $\mathbf{w}^{(i)}$ を更新する。従って、CW の最適化問題は、下のように定式化される。

$$\begin{aligned}(\boldsymbol{\mu}^{(i+1)}, \Sigma^{(i+1)}) = \arg\min_{\mu, \Sigma} \; & D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \parallel \mathcal{N}(\boldsymbol{\mu}^{(i)}, \Sigma^{(i)})) \\ s.t. \; & Pr_{\mathbf{w}}[y^{(i)}(\mathbf{w} \cdot x^{(i)}) \geq 0] \geq \eta\end{aligned} \tag{2}$$

2 値分類問題の場合は、データがある 1 つのクラスに所属するかしないかを調べるだけで、データが所属するクラスを予測することが出来る。そのため、前章で導入した特徴ベクトル $f(\mathbf{x}, y)$ を使用しなくても最適化問題が定式化出来る。

そして、Crammer et al. [2] によって、この最適化問題は閉じた式で解ける事が示されている。更新式は、以下の式で表される。

$$\boldsymbol{\mu}^{(i+1)} = \boldsymbol{\mu}^{(i)} + \alpha^{(i)} \Sigma^{(i)} \mathbf{x}^{(i)} \tag{3}$$

$$\Sigma^{(i+1)} = \left((\Sigma^{(i)})^{-1} + \beta^{(i)} \mathbf{x}^{(i)} (\mathbf{x}^{(i)})^T \right)^{-1} \tag{4}$$

ガウス分布の共分散行列 Σ はパラメータの不確実性を表しており、出現頻度が低く更新回数の少ないパラメータほど、更新時に対応するパラメータが大きく変化する学習方法を実現している。

さらに、Crammer et al. [3] によって、CW の多クラス分類問題への拡張が行われている。多クラス分類問題の場合も 2 値分類の場合と同様に、入力ベクトル $\mathbf{x}^{(i)}$ が与えられたとき、重みベクトル $\mathbf{w}^{(i)}$ を用いて最適なクラスを予測する。しかし、多クラスの場合は 2 値分類の場合と異なり、ある 1 つのクラスへの所属の有無を調べるだけではクラスを予測することが出来ない。従って、受け取ったデータの所属するクラスを予測するためには、$\arg\max_y \mathbf{w} \cdot f(\mathbf{x}, y)$ を求める必要がある。

しかし、$Pr[y^{(i)} = \arg\max_y \mathbf{w} \cdot f(\mathbf{x}^{(i)}, y)] \geq \eta$ の条件から与えられる最適化問題は凸性を持たないため [3]、更新式を閉じた形で書くことが難しい。よって最適化問題の条件式を以下の式に緩和する。

$$Pr[\mathbf{w} \cdot f(\mathbf{x}^{(i)}, y^{(i)}) \geq \mathbf{w} \cdot f(\mathbf{x}^{(i)}, k)] \geq \eta \quad (\forall k \neq y^{(i)}) \tag{5}$$

上式から分かるように、正解クラスとある不正解クラスの 2 値分類問題を考えたときに、正しく識別出来る確率が η 以上になる制約を、全てのクラスに対して満たすことが、最適化問題の条件式になっている。

しかし、上記の手法では、共分散行列のパラメータ数が $NK \times NK$ 次元となり、非常に次元数の大きな行列となるため、上の最適化問題を厳密に適用させることはなお難しい。また、著者らは全てのクラスについて条件を成立させるように最適化問題を解くと、過学習が発生することがある事を指摘している。そのため、多クラス分類問題に対する CW を提案した Crammer et al. [3] では、

- 重みベクトルの行分散行列は対角項しか考慮しない
- 上位数クラスのみに対して最適化問題の条件とする

と、近似的な手法が取られている。

本研究では、後の実験の際、正解クラスとその他のクラスで最上位のクラスのみを最適化問題の条件とする Single Constraint のアルゴリズムと、全てのクラスについて条件式を設定する $k = \infty$ のアルゴリズムを既存手法として扱っている。本来は、上位数個のクラスに対して条件式を設定したアルゴリズムも対象とすべきであるが、本研究では割愛する。

また、CW 以外にも識別問題に対するオンライン学習アルゴリズムの先行研究として、Perceptron [6]、Passive-Aggressive(PA) [5] などの手法が提案されている。

4 提案手法

本章では、CW を多クラス分類問題に拡張する手法を提案する。本章では、提案手法となるアルゴリズムが導出される手順に従って、3 つの節に分けて手法の説明を行う。

初めに、4.1 節では、3 章で述べた多クラス分類問題に対するアルゴリズムの問題点を解決するため、重み行列上にガウス分布を導入する手法を提案する。この手法を用いることで、重みベクトルに直接ガウス分布を導入する場合に比べ、パラメータ数を減少させることが出来、次元数が大きくなる問題を緩和させることが出来る。

次に、4.2 節では重み行列にガウス分布を導入した場合の最適化問題を定式化し、パラメータの更新式を求める。

最後に、4.3 節ではサポートクラスと呼ばれる概念を適用する。サポートクラスを導入することで、4.2 節で求めた最適化問題の更新式を、厳密かつ効率的に解く方法を提案する。

4.1 重み行列上へのガウス分布の導入

はじめに、3 章で紹介した重みベクトル $\mathbf{w}$ を重み行列 W の形に直す。

$$W = \begin{pmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_K \end{pmatrix}$$

ここで、$\mathbf{w}_k$ は D 次元のベクトルである。

次に、重み行列上にガウス分布を導入する。

$$W \sim \mathcal{N}_{D\times K}\left(\begin{pmatrix}\boldsymbol{\mu}_1 & \boldsymbol{\mu}_2 & \dots & \boldsymbol{\mu}_K\end{pmatrix}, T, \Sigma\right) \tag{6}$$

ここで、重み行列 W 上のガウス分布は、各クラスの各特徴ベクトルの平均の値を示す行列 $(\boldsymbol{\mu}_1\ \boldsymbol{\mu}_2\ \dots\ \boldsymbol{\mu}_K)$、クラス間の関係を表す共分散行列 $T \in \Re^{K\times K}$ と、特徴ベクトルの共分散行列 $\Sigma \in \Re^{D\times D}$ の 3 つの行列を用いて、上記のように表すことが出来る。

そして、この重み行列 W をベクトル化すると、導出される重みベクトル w は、クロネッカー積 $\otimes$ を用いて、

$$vec(W) = \mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma \otimes T) \tag{7}$$

と表現することが出来る。

ここで、クロネッカー積とは、ある $m \times n$ の行列 A と、ある $o \times p$ の行列 B が存在したときに、

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{pmatrix}$$

の形の行列を与える演算のことである。

さらに、各クラスに対応する重みベクトルの平均は、$\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, ..., \boldsymbol{\mu}_K^T)^T$ とおいている。

重みベクトルの共分散行列を 2 つの共分散行列のクロネッカー積で表すことで、共分散行列のパラメータ数が $DK \times DK$ から $(D \times D) + (K \times K)$ に変化する。従って、D や K が大きい場合には、従来の手法に比べ、パラメータ数を大幅に減少させることが出来る。さらに、クラス間の関係を表す行列 T は、学習前にある程度値を予測する事が可能なため、定数とする事も可能である。

本研究では簡単のため、T は単位行列とする。従って、パラメータ数は $D \times D$ まで減少する。

4.2 最適化問題と更新式の記述

重み行列上にガウス分布を適用した場合の最適化問題を定式化すると、

$$\begin{aligned}(\boldsymbol{\mu}^{(i+1)}, \Sigma^{(i+1)}) = \arg\min_{\mu,\Sigma}\ & D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma \otimes I) \parallel \mathcal{N}(\boldsymbol{\mu}^{(i)}, \Sigma^{(i)} \otimes I)) \\ & s.t. Pr_{\mathbf{w}}[\mathbf{w} \cdot f(\mathbf{x}^{(i)}, y^{(i)}) \geq \mathbf{w} \cdot f(\mathbf{x}^{(i)}, k)] \geq \eta \quad (\forall k \neq y^{(i)})\end{aligned} \tag{8}$$

となる。

初めに、制約式の式変形を行う。$\mathbf{w} \cdot f(\mathbf{x}^{(i)}, y^{(i)}) \sim \mathcal{N}\left(\boldsymbol{\mu}_{y^{(i)}} \cdot \mathbf{x}^{(i)}, (\mathbf{x}^{(i)})^T \Sigma \mathbf{x}^{(i)}\right)$ より、

$$\mathbf{w} \cdot f(\mathbf{x}^{(i)}, y^{(i)}) - \mathbf{w} \cdot f(\mathbf{x}^{(i)}, k) \sim \mathcal{N}\left((\boldsymbol{\mu}_{y^{(i)}} - \boldsymbol{\mu}_k) \cdot \mathbf{x}, 2(\mathbf{x}^{(i)})^T \Sigma \mathbf{x}^{(i)})\right)$$

と表すことが出来る。従って、$\mathbf{w} \cdot f(\mathbf{x}^{(i)}, y^{(i)}) - \mathbf{w} \cdot f(\mathbf{x}^{(i)}, k) = M$ とおくと

$$\begin{aligned} & Pr[M \geq 0] \geq \eta \\ & \quad = Pr\left[\frac{M - (\boldsymbol{\mu}_{y^{(i)}} - \boldsymbol{\mu}_k) \cdot \mathbf{x}}{2(\mathbf{x}^{(i)})^T \Sigma \mathbf{x}^{(i)}} \geq -\frac{(\boldsymbol{\mu}_{y^{(i)}} - \boldsymbol{\mu}_k) \cdot \mathbf{x}}{2(\mathbf{x}^{(i)})^T \Sigma \mathbf{x}^{(i)}}\right] \geq \eta \end{aligned} \tag{9}$$

となる。

$\frac{M-(\boldsymbol{\mu}_{y^{(i)}}-\boldsymbol{\mu}_k)\cdot\mathbf{x}}{2(\mathbf{x}^{(i)})^T\Sigma\mathbf{x}^{(i)}}$ は標準正規分布なので、標準正規分布の累積密度関数 ϕ の逆関数を用いて、

$$(\boldsymbol{\mu}_{y^{(i)}}-\boldsymbol{\mu}_k)\cdot\mathbf{x}\geq\phi^{-1}(\eta)\sqrt{2(\mathbf{x}^{(i)})^T\Sigma\mathbf{x}^{(i)}}\quad(\forall k\neq y^{(i)}) \tag{10}$$

と整理することが出来る。

一方、目的関数は、

$$\begin{aligned}&D_{KL}(\mathcal{N}(\boldsymbol{\mu},\Sigma\otimes I)\parallel\mathcal{N}(\boldsymbol{\mu}^{(i)},\Sigma^{(i)}\otimes I))\\&\quad=\frac{K}{2}log\frac{|\Sigma^{(i)}|}{|\Sigma|}+\frac{K}{2}tr((\Sigma^{(i)})^{-1}\Sigma)+\sum_{k=1}^{K}\frac{1}{2}(\boldsymbol{\mu}_k^{(i)}-\boldsymbol{\mu}_k)^T(\Sigma^{(i)})^{-1}(\boldsymbol{\mu}_k^{(i)}-\boldsymbol{\mu}_k)\end{aligned} \tag{11}$$

と書き直すことが出来る。Σ は正定値行列なので、対角行列 Q と Σ の全ての固有値の $\frac{1}{2}$ 乗を対角項に並べた行列を用いて、$\Sigma=\Upsilon^2$, $\Upsilon=Q^Tdiag(\lambda_1^{\frac{1}{2}},\lambda_2^{\frac{1}{2}},...,\lambda_K^{\frac{1}{2}})Q$ と書き表すことが出来る。制約条件 (10) は、Υ,Σ について凸関数とすることが出来る [3] ことが示されている。

従って、上記の目的関数 (11) と制約条件 (10) で定式化された最適化問題は、KKT 条件を満たす解が最適解となることが保証される。KKT 条件は以下のように定式化される。

$$\begin{aligned}L=&\frac{K}{2}log\frac{|\Upsilon^{(i)}|^2}{|\Upsilon|^2}+\frac{K}{2}tr((\Upsilon^{(i)})^{-2}\Upsilon^2)+\sum_{k=1}^{K}\frac{1}{2}(\boldsymbol{\mu}_k^{(i)}-\boldsymbol{\mu}_k)^T(\Upsilon^{(i)})^{-2}(\boldsymbol{\mu}_k^{(i)}-\boldsymbol{\mu}_k) &(12)\\&+\sum_{k\neq y^{(i)}}\alpha_k^{(i)}\left(\Phi\sqrt{2(\mathbf{x}^{(i)})^T\Upsilon^2\mathbf{x}^{(i)}}-(\boldsymbol{\mu}_{y^{(i)}}-\boldsymbol{\mu}_k)\cdot\mathbf{x}^{(i)}\right)\\&\alpha_k^{(i)}\left(\phi^{-1}(\eta)\sqrt{2(\mathbf{x}^{(i)})^T\Upsilon^2\mathbf{x}^{(i)}}-(\boldsymbol{\mu}_{y^{(i)}}-\boldsymbol{\mu}_k)\cdot\mathbf{x}^{(i)}\right)=0 \qquad(\forall k\neq y^{(i)}) &(13)\\&\alpha_k^{(i)}\geq 0 \qquad(\forall k\neq y^{(i)}) &(14)\\&\left(\phi^{-1}(\eta)\sqrt{2(\mathbf{x}^{(i)})^T\Upsilon^2\mathbf{x}^{(i)}}-(\boldsymbol{\mu}_{y^{(i)}}-\boldsymbol{\mu}_k)\cdot\mathbf{x}^{(i)}\right)\leq 0 \qquad(\forall k\neq y^{(i)}) &(15)\end{aligned}$$

ここで、表記を単純にするため $\phi^{-1}(\eta)=\Phi$ としている。

この最適化問題を解くため、まず μ で偏微分すると、

$$\begin{aligned}\boldsymbol{\mu}_k=&\quad\boldsymbol{\mu}_k^{(i)}-\alpha_k^{(i)}\Sigma^{(i)}\mathbf{x}^{(i)} \qquad(\forall k\neq y^{(i)}) &(16)\\\boldsymbol{\mu}_k=&\boldsymbol{\mu}_k^{(i)}+\sum_{k\neq y^{(i)}}\alpha_k^{(i)}\Sigma^{(i)}\mathbf{x}^{(i)} \qquad(k=y^{(i)}) &(17)\end{aligned}$$

が導出され、μ の更新式が記述できる。

次に、Υ で偏微分すると、

$$\Upsilon^{-2}=\Upsilon_i^{-2}+\sum_{k\neq y^{(i)}}\alpha_k^{(i)}\Phi\frac{2\mathbf{x}^{(i)}(\mathbf{x}^{(i)})^T}{K\sqrt{2(\mathbf{x}^{(i)})^T\Upsilon^2\mathbf{x}^{(i)}}} \tag{18}$$

が導出される。(18) 式は $Woodbury\ identity$ を用いると、

$$\Sigma=\Sigma^{(i)}-\Sigma^{(i)}x^{(i)}\left(\frac{\sqrt{2}(\sum_{k\neq y^{(i)}}\alpha_k^{(i)})\Phi}{K\sqrt{(\mathbf{x}^{(i)})^T\Sigma\mathbf{x}^{(i)}}+\sqrt{2}(\sum_{k\neq y^{(i)}}\alpha_k^{(i)})\Phi(\mathbf{x}^{(i)})^T\Sigma^{(i)}\mathbf{x}^{(i)}}\right)(\mathbf{x}^{(i)})^T\Sigma^{(i)} \tag{19}$$

となり、Σ の更新式が導かれる。ここで、表記の簡単化のため、$u^{(i)} = (\mathbf{x}^{(i)})^T \Sigma \mathbf{x}^{(i)}, v^{(i)} = (\mathbf{x}^{(i)})^T \Sigma^{(i)} \mathbf{x}^{(i)}$ とおいている。

さらに、$\sqrt{u^{(i)}}$ を求めるため、(19) 式の左右から $(\mathbf{x}^{(i)})^T, \mathbf{x}^{(i)}$ を掛ける。すると $\sqrt{u^{(i)}}$ は、以下のように導出される。

$$\sqrt{u^{(i)}} = \frac{-\sqrt{2}(\sum_{k \neq y^{(i)}} \alpha_k^{(i)})\Phi v^{(i)} + \sqrt{2\Phi^2 (v^{(i)})^2 (\sum_{k \neq y^{(i)}} \alpha_k^{(i)})^2 + 4K^2 v^{(i)}}}{2K} \tag{20}$$

この式変形によって、あとは $(\sum_{k \neq y^{(i)}} \alpha_k^{(i)})$ を求めることが出来れば、μ, Σ の更新を閉じた形で記述することが可能になる。

4.3 サポートクラスによる厳密解の導出

本研究では、$\alpha_k^{(i)}$ の値を求めるため、サポートクラス [4] と呼ばれる概念を導入する。サポートクラスとは、$(\boldsymbol{\mu}_{y^{(i)}} - \boldsymbol{\mu}_k) \cdot \mathbf{x}^{(i)} = \Phi \sqrt{2u^{(i)}}$ となるクラス k を指す。つまり、制約条件が有効に働くクラスをサポートクラスと呼ぶ。

KKT 条件より、サポートクラスに対応する Lagrange 乗数は $\alpha_k^{(i)} > 0$ となり、一方、サポートクラス以外のクラスに対応する $\alpha_k^{(i)}$ は 0 になる。ここで、サポートクラスとなるクラスの集合をサポートクラス集合 S とおく。これから、サポートクラス集合 S を用いて、具体的な $\alpha_k^{(i)}$ の値を求める。証明の手順は、以下の通りである。

1. はじめに、S に含まれるクラスが分かっていると仮定した上で、$\alpha_k^{(i)}$ の具体的な値を求める。
2. 次に、S に含まれるクラスを導出する。

はじめに、サポートクラス集合 S に含まれるクラスが分かっていると仮定した上で、S の中のあるクラス k に対応する $\alpha_k^{(i)}$ の値を導出する。クラス k に対応する制約式は、$(\boldsymbol{\mu}_{y^{(i)}} - \boldsymbol{\mu}_k) \cdot \mathbf{x}^{(i)} = \Phi \sqrt{2u^{(i)}}$ である。この式に (16),(17),(20) 式を代入して整理すると、

$$\left(\sum_{s \neq y^{(i)}} \alpha_s^{(i)} \Sigma^{(i)} \mathbf{x}^{(i)} + \alpha_k^{(i)} \Sigma^{(i)} \mathbf{x}^{(i)} + \mu_{y^{(i)}}^{(i)} - \mu_k^{(i)} \right) \cdot \mathbf{x}^{(i)} = \Phi \frac{-(\sum_{s \neq y^{(i)}} \alpha_s^{(i)})\Phi v^{(i)} + \sqrt{\Phi^2 (v^{(i)})^2 (\sum_{s \neq y^{(i)}} \alpha_s^{(i)})^2 + 2K^2 v^{(i)}}}{K} \tag{21}$$

となる。(21) 式を S に属する全てのクラスに関して総和をとる事で、$\sum_{s \neq y^{(i)}} \alpha_s^{(i)}$ を導出できる。S に属するクラスの数を $|S|$ で表し、また式の簡略化のため、$\sum_{s \neq y^{(i)}} \alpha_s^{(i)} = A$, $(|S| + 1)Kv^{(i)} + |S|\Phi^2 v^{(i)} = B$ $(\mu_{y^{(i)}}^{(i)} - \mu_k^{(i)}) \cdot \mathbf{x}^{(i)} = l_k^{(i)}$ とおくと、

$$A^2 \left(B^2 - |S|^2 \Phi^4 (v^{(i)})^2 \right) + 2AKB \sum_{k \in S} l_k^{(i)} + K^2 \left(\left(\sum_{k \in S} l_k^{(i)} \right)^2 - 2|S|^2 \Phi^2 v^{(i)} \right) = 0$$

と式を整理することが出来るため、二次方程式の解の公式より A を求めることが出来る。

$$A = \frac{-KB\sum_{k\in S} l_k^{(i)} + |S|K\Phi\sqrt{\left(\Phi^2(v^{(i)})^2\right)\left(\sum_{k\in S} l_k^{(i)}\right)^2 + 2v^{(i)}\left(B^2 - |S|^2\Phi^4(v^{(i)})^2\right)}}{\left(B^2 - |S|^2\Phi^4(v^{(i)})^2\right)} \tag{22}$$

さらに、(22) 式を (21) 式に代入すれば、サポートクラス k に対応する $\alpha_k^{(i)}$ の値も求めることが出来る。

$$\alpha_k^{(i)} = \frac{-l_k^{(i)} + \Phi\sqrt{2u^{(i)}}}{v^{(i)}} - A \tag{23}$$

以上で求めた $\sum_{s\neq y^{(i)}} \alpha_s^{(i)}$,$\alpha_k^{(i)}$ を (16),(17),(19) 式に代入すれば、更新式を閉じた形で記述可能になる。

次に、サポートクラス集合 S に所属するクラスを求める。S に含まれないクラス k は、(16) 式の $\alpha_k^{(i)}$ が 0 となるため、(16),(17) 式を (15) 式に代入すると、$\left(\Phi\sqrt{2u^{(i)}} - Av^{(i)} - l_k^{(i)}\right) \leq 0$ より、

$$\frac{-l_k^{(i)} + \Phi\sqrt{2u^{(i)}}}{v^{(i)}} - A \leq 0 \tag{24}$$

が導出される。また、S に所属するクラスは $\alpha_k^{(i)} > 0$ となるため、(23) 式より、

$$\frac{-l_k^{(i)} + \Phi\sqrt{2u^{(i)}}}{v^{(i)}} - A > 0 \tag{25}$$

となる。従って、あるクラス A より $l_k^{(i)}$ の値が大きいクラス k がサポートクラスであれば、クラス A はサポートクラスであり、あるクラス A より $l_k^{(i)}$ の値が小さいクラス k がサポートクラスでないならば、クラス A もサポートクラスではないことが示される。従って、S に含まれるクラスは、下のアルゴリズムによって求めることが出来る。

1. $l_k^{(i)}$ が小さい順にクラスを並べ、はじめに $l_k^{(i)}$ が一番小さいクラスを選ぶ。
2. 選択したクラスより $l_k^{(i)}$ が小さいクラス全て（選択したクラスを含む）がサポートクラスであると仮定して、選択したクラスの $\alpha_k^{(i)}$ を求める。
3. $\alpha_k^{(i)} > 0$ であれば、$l_k^{(i)}$ が次に大きいクラスを選択し、2. に戻る。
4. $\alpha_k^{(i)} \leq 0$ であれば、選択したクラスより $l_k^{(i)}$ が大きいクラスはサポートクラスにならないため、サポートクラス集合は選択したクラスより $l_k^{(i)}$ が小さいクラス全てに決まる。

以上の議論により、S に含まれるクラスが一意に決定出来る事を証明した。**Algorithm 1** は、上記のアルゴリズムを整理したものである。

5 実験

行列上のガウス分布を用いた提案手法[*1]を、実在の多クラス識別問題に適用することで、性能評価を行った。

実験の手法について説明する。評価のためのデータセットには 20 NewsGroups[*2] 、Reuters-21578[*3]、USPS[*4] 、を用いた。

[*1] SCCW と表記する
[*2] http://mlg.ucd.ie/datasets
[*3] http://www.daviddlewis.com/resources/testcollections/reuters21578
[*4] http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets

Algorithm 1 SupportClassCW algorithm

入力:ハイパーパラメータ η

初期値:$\boldsymbol{\mu}^{(1)} = \mathbf{0}$, $\Sigma^{(1)} = I$, $\Phi = \phi^{-1}(\eta)$,

For $i = 1, 2, ...$

特徴ベクトル $\mathbf{x}^{(i)}$ と正解ラベル $y^{(i)}$ を受け取る

正解クラスを除く全てのクラスについて $\boldsymbol{\mu}_k^{(i)} \cdot \mathbf{x}^{(i)}$ を計算し、クラスをこの内積が大きい順に並べる

（内積の大きいクラスから、番号 j を $1, 2, ...$ と割り振る）

For $j = 1, 2, ...$

j 以下の全てのクラスがサポートクラスであると仮定して、$\alpha_j^{(i)} = \frac{-l_j^{(i)} + \Phi\sqrt{2u^{(i)}}}{v^{(i)}} - A$ を計算する

(A は (22) 式を用いて導出する)

If $\alpha_j^{(i)} \leq 0$ **break**

$\alpha_k^{(i)} > 0$ となる $1 \sim k$ のクラスをサポートクラス S として、パラメータを更新する

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

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

$$\Sigma^{(i+1)} = \Sigma^{(i)} - \left(\frac{\sqrt{2}(\sum_{k \in S} \alpha_k^{(i)})\Phi}{K\sqrt{u^{(i)}} + \sqrt{2}(\sum_{k \in S} \alpha_k^{(i)})\Phi (\mathbf{x}^{(i)})^T \Sigma^{(i)} \mathbf{x}^{(i)}} \right) \Sigma^{(i)} \mathbf{x}^{(i)} (\mathbf{x}^{(i)})^T \Sigma^{(i)} \quad \text{[MCCW]}$$

$$\Sigma^{(i+1)} = \Sigma^{(i)} - \left(\frac{\sqrt{2}(\sum_{k \in S} \alpha_k^{(i)})\Phi}{K\sqrt{u^{(i)}} + \sqrt{2}(\sum_{k \in S} \alpha_k^{(i)})\Phi (\mathbf{x}^{(i)})^T \Sigma^{(i)} \mathbf{x}^{(i)}} \right) diag\left(\Sigma^{(i)} \mathbf{x}^{(i)} (\mathbf{x}^{(i)})^T \Sigma^{(i)} \right) \quad \text{[MCCWD]}$$

出力:重みパラメータ $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma \otimes T)$

20 NewsGroups は約 20000 件、20 カテゴリのニュース記事からなるデータセットである。本研究では、20 NewsGroups の中でも、6 つのサブセットを用いた。各サブセットは、記事の精度やデータ数に従ってに分けられている。

1 文字目のアルファベットは'o' は'overlapped'、's' は'separeted' を意味している。'o' の方が精度が低くなっている。2 文字目のアルファベットは、クラスのデータ分布を表している。'b' は'balanced' を意味し、クラス間でデータ数が均等になっている。's' は'small' を意味し、特定のクラスのデータ数が小さくなっている。'l' は'large' を意味し、特定のクラスのデータ数が大きくなっている。次の数字はクラス数を意味している。データは bag-of-words で与えられており、特徴次元数が非常に大きいデータセットである。

Reuters-21578(reut20) もニュース記事からなるデータセットである。本研究では、このコーパスから 20 クラスの分類問題を作成し、使用した。

USPS は手書き数字認識問題からなるデータセットである。

以上で説明したとおり、データセットによってデータ数・特徴次元数・クラス数が異なる。表 1 がデータセットの概要である。

本研究では、提案手法として SCCW と共分散行列の対角項のみを更新する SCCWD を用いた。さらに、オンライン学習による従来の手法として Passive-Aggressive [5](PA,PA-I,PA-II) と、Multi-class CW [3] の

	データ数	特徴次元数	クラス数
ol-7-1	2,586	9,605	7
ol-8-1	2,388	9,971	8
ob-7-1	3,500	12,878	7
ob-8-1	4,000	13,890	8
sb-7-1	3,500	13,997	7
sb-8-1	4,000	16,282	8
reut20	7,800	34,488	20
USPS	7,291	256	10
News20	15,935	60,345	20

表 1　データセットの概要

	PA	PA-I	PA-II	CW $k=1$	CW $k=\infty$	SCCW	SCCWD
ol-7-1	87.2281	88.9445	88.6518	**94.1373**	91.1647	93.0910	93.0491
ol-8-1	87.8975	89.6768	89.6765	**94.8182**	91.9952	94.0075	94.0077
ob-7-1	88.9429	91.0571	90.7714	**95.6286**	91.8286	94.8571	94.8571
ob-8-1	90.3000	90.3750	90.4500	**94.8182**	91.9952	94.6250	94.6500
sb-7-1	92.2286	92.8286	92.6857	**96.4857**	95.3714	96.1429	96.1429
sb-8-1	92.1750	92.5000	92.7750	**96.9750**	94.8000	96.2500	96.2500
reut20	94.3590	94.3590	94.1923	95.8205	93.9231	**96.5128**	96.3590
USPS	90.0424	89.6175	90.4130	92.4013	83.5274	**93.7318**	93.5126
news20	75.7579	78.0483	78.0483	**84.7568**	76.3162	83.3134	83.2883

表 2　各分類手法の正識別率（%） 反復回数:3 回

$k=1$(Single Constraint), $k=\infty$ Sequantial に対して比較を行った[*5]。

各分類手法について、3 回教師データを反復計算をして学習をした後に、訓練データを用いて性能を評価している。また、データセットを 10 分割して Cross-Validation を行う。分析結果では、Cross-Validation による 10 回の分類結果を平均した数値を分類精度として用いる。

分析の結果を表 2 に示す。各データセットについて、一番分類精度の高い結果を太字で示している。表 2 から、reut20、USPS に対しては、今回の提案手法が従来手法よりも高い分類性能を示している事が確認出来る。一方、news20 とそのサブセットでは、CW に比べて分類精度の改善は見られない。

また、CW と SCCW,SCCWD の reut20 における反復回数毎の分類精度の変化について、図 1 で示す。SCCW と CW を比較すると、反復回数が少ない時は CW と SCCW の間に大きな違いは見られないが、反復回数が増えるにつれ、SCCW の精度が CW の精度を上回っている事が分かる。

[*5] 本章では、表記の簡単化のため Multi-Class CW を CW と表記する

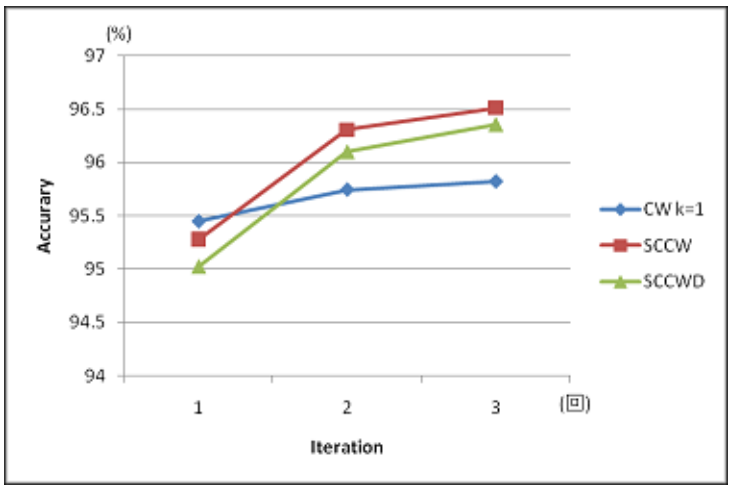

図 1　反復回数毎の分類精度

6　まとめ

本研究では、多クラス分類問題に対する CW の拡張の方法として、行列上にガウス分布を導入する手法を提案した。この手法を用いることで、更新に必要なパラメータ数を減少させることに成功した。さらに、本手法を実データに基づく多クラス分類問題に適用することで、その性能を評価した。

しかし、本研究ではクラス間の関係を表す T を単位行列と置いた事で、各クラスのデータ数の違いなどに対応したパラメータ更新を行うことが出来ていない。また、実験でも、CW の特定の手法・特定のデータセットに対する性能比較を行ったのみである。

そこで、今後の課題として、

- クラス間の関係を表す行列 T を更新する手法
  - 観測した各クラスのデータ数に応じて、T の対角成分を更新する手法
  - クラス間の近さを表現する手法
- 提案手法の Mistake Bound Analysis
- 他のデータセット／提案手法に対する評価実験

の検討を行っていく予定である。

参考文献

[1] Dredze, M., Crammer, K., & Pereira, F. ”Confidence-weighted linear classification” International Conference on Machine Learning(ICML), 2008

[2] Crammer, K., Dredze, M., & Pereira, F. ”Exact Confidence-weighted Learning” Advances in Neural Information Processing System, 2008

[3] Crammer, K., Dredze, M., & Pereira, F. ”Multi-Class Confidence Weighted Algorithms” Empirical

Methods in Natural Language Processing, 2009

[4] Shin, M., Nobuyuki, K., Kazuhiro, Y., Takashi, N., & Hiroshi, N. "Exact Passive-Aggressive Algorithm for Multiclass Classification Using Support Class" SIAM International Conference on Data Mining, 2010

[5] Crammer, K., Dekel, O., & Keshet, J. "Online passive-aggressive algorithms" Journal of Machine Learning Research, 2006

[6] Novikoff, A. B. J., "On convergence proofs on perceptrons" In Proc. of Symposium on the Mathematical Theory of Automata, 1962