1G2-4

識別における汎化中立性の保証

Generalization Neutrality Guarantee for Classification

福地 一斗 *1　佐久間 淳 *1*2
Kazuto Fukuchi　Jun Sakuma

*1筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻
Dept. of Computer Science, Graduate school of SIE, University of Tsukuba

*2科学技術新興機構 CREST
Japan Science and Technology Agency CREST

In this paper, we introduce a novel framework of empirical risk minimization (ERM), neutralized ERM (NERM) that guarantees that the prediction by the target hypothesis of NERM does not cause discrimination, unfairness treatment or biased view with respect to the viewpoint hypothesis. We provides the theoretical analysis of the generalization neutrality bound of NERM. Furthermore, we derive a max-margin algorithm for linear classification, neutral support vector machine (SVM) that follows the NERM principle. We show the neutral SVM improves the classification accuracy without sacrificing neutrality.

1. はじめに

Empirical Risk Minimization (ERM) は，**入力** x と**目標** y の集合に対する経験損失が最小となる仮説 f を獲得することで教師付き学習を行う枠組みである．本稿では，ERM に対して新たに**視点仮説** g を導入し，視点仮説に対する中立化について述べる．仮説 f は入力 x に対する目標の予測 $y = f(x)$ を与える関数であり，視点仮説 g は与えられた入力 x に対する視点の予測 $v = g(x)$ を与える関数である．f と g を区別するため，f を**目標仮説**と呼ぶ．目標仮説が視点仮説に対して中立であるとは目標 $f(x)$ と視点 $g(x)$ との間の相関が小さい状態のことを指し，中立化とは教師あり学習において予測の精度を保持したまま，与えられた g に対して中立な目標仮説 f を得るための学習法である．本稿では，ERM を基にした中立化の新しい枠組みである *Neutralized ERM* (NERM) を提案する．

中立化が解決する問題の一つとして，*filter bubble* [Pariser 11] があげられる．例えば，ユーザの興味に応じた記事配信システムを考える．このとき，入力 x としてアクセスログなどを収集し，目標 y である記事がユーザの好みかどうかを目標仮説 $y = f(x)$ によって予測する．あるユーザは世論を 2 分するような政策に偏った意見を持っており，視点仮説 $v = g(x)$ によってどちらの意見なのか予測できたとする．もし，$f(x)$ と $g(x)$ が強く相関していると，片方の意見に関する記事ばかり推薦していることになり，政策に偏見を与えかねない．偏見を排除するためには，目標仮説と視点仮説の出力 $f(x), g(x)$ が互いに相関しないようにする必要がある．

記事配信システムは，過去のユーザの記事の好みに関するデータを用いて学習を行うが，推薦はまだ読まれていない記事に対してする必要がある．従って，目標仮説から偏見を排除するためには，未知の記事に対して中立である必要がある．このように，教師あり学習における中立性は，未知の入力 x に対応する目標 $f(x)$ と視点 $g(x)$ の相関性によって測られる．これは，教師あり学習において分類器の精度を測る汎化損失と似たような基準であることから，未知の事例に対する中立性の性能を汎化中立性と呼ぶ．教師あり学習における中立化の目的は，汎化中立性が保証されると同時に汎化損失を最小となる目標仮説を獲得することである．汎化中立性と汎化損失の間のよいトレードオフを得ることが，中立化における課題となる．

本稿では，分類問題において中立化を行う学習アルゴリズムの枠組みとして NERM を提案する．NERM は，出力 y が 2 値である ERM において，2 値である視点 v に対して中立化を行うアルゴリズムを構築できる．NERM は，ERM に対して中立性が低いことに対する罰則項を加えた最適化問題として定式化される．最適化問題の目的関数は，パラメータによって分類と中立性の性能のトレードオフを制御することができる．NERM の最適化問題は凸となるため，NERM によって構築された学習手法は大域的最適解を保証できる．

中立化を行う方法としてヒューリスティックな方法 [Calders 10] や最適化を基にした手法 [Kamishima 12], [Zemel 13], [Fukuchi 13] があり，どの手法も与えられた事例について経験的に計算される中立性の評価量を基にして中立化を行っているため，未知事例に対する中立性の保証は無い．しかし，汎化中立性を保証するための，理論的な解析は行われていない．本稿では，NERM の枠組みにおける汎化中立性の確率的バウンドに関する理論的解析を行う．

NERM の枠組みを用いた適用例として，2 値の線形分類器として高い性能が示されている *Support Vector Machine* [Vapnik 98] において中立化を行う．提案する中立 SVM は，双対問題を導くことによってカーネル化を行うことができる．カーネル化を行うことによって入力の非線形な特徴を用いて分類を行うことができ，分類の精度と中立性の間の高いトレードオフが実現できることが期待できる．

2. 汎化中立性リスクと経験中立性リスク

X, Y をそれぞれ入力と目標の空間，$D_n = \{(x_i, y_i)\}_{i=1}^n \in Z^n$ $(Z = X \times Y)$ を $(Z, \mathcal{Z})$ 上の未知の確率測度 ρ から i.i.i. に生成された事例集合であるとし，入力 X から 2 値の目標 $Y = \{-1, 1\}$ の予測を行う教師あり学習の問題を考える．教師あり学習では，仮説集合 $f \in \mathcal{F}$ から目標 y のよい予測が可能な目標仮説 $f : X \to \mathbb{R}$ を，与えられた事例集合を用いて選択

連絡先: 福地 一斗，筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻，茨城県つくば市天王台 1-1-1，029-853-3826，kazuto@mdl.cs.tsukuba.ac.jp

する．目標 y は $f(x) > 0$ であれば $y = 1$，それ以外であれば $y = -1$ と予測するとし，目標仮説 f による目標 y の分類結果は $\mathrm{sgn} \circ f$ となる．教師あり学習の目的は，真の目標 y と目標の予測 $f(x)$ の間の損失関数を $\ell : Y \times \mathbb{R} \to \mathbb{R}^+$ としたとき，汎化損失 $R(f) = \int \ell(y, f(x))d\rho$ が最小となる目標仮説 $f^* \in \mathcal{F}$ を獲得することである．しかし，確率測度 ρ は未知であるため，汎化損失を直接評価することはできない．Empirical Risk Minimization (ERM) は，汎化損失に代わって，与えらた事例集合 D_n に関する経験損失 $R_n(f) = \sum_{i=1}^n \ell(y_i, f(x_i))/n$ を最小とする目標仮説を学習結果とする枠組みである．また，仮説の複雑さを抑制するために，正則化項 $\Omega : \mathcal{F} \to \mathbb{R}^+$ を正則化パラメータ $\lambda \geq 0$ とともに加えた目的関数を最小化する枠組みがあり，Regularized ERM (RERM) と呼ばれる．

本稿では，ERM における教師あり学習の枠組みをもとに，与えられた視点仮説に対する中立化を行う枠組みである Neutralized ERM (NERM) を定義する．NERM では，与えられた事例集合 D_n 内の事例ではなく，未知の事例も含めたすべての事例における中立性の評価量として汎化中立性リスクを導入する．また，本章では中立性リスクに基づいた最適化を行うために，中立性リスクの凸緩和についても述べる．

2.1 +1/-1 汎化中立性リスク

ERM における教師あり学習の枠組みに基づいて中立化を行うために，中立化を行う対象である**視点仮説**を導入する．可測関数 $g : X \to \mathbb{R}$ を**視点仮説**と呼び，視点仮説 g によって得られる予測 $v = g(x)$ を**視点**と呼ぶ．目標仮説 f と視点仮説 g はどちらも 2 値分類を行うとし，f もしくは g はそれぞれ $\mathrm{sgn} \circ f$，$\mathrm{sgn} \circ g$ によって予測する．目標仮説 f が視点仮説 g に対して中立であるとは，f による予測 $\mathrm{sgn} \circ f$ と g による予測 $\mathrm{sgn} \circ g$ が確率測度 ρ のもとで互いに相関しないことを指す．多くの事例について $f(x)g(x) > 0$ が成り立つならば，目標 $\mathrm{sgn} \circ f$ と視点 $\mathrm{sgn} \circ g$ は ρ に基づく入力 x についてほぼ同じ出力であり，互いに相関していると考えられる．また，多くの事例について $f(x)g(x) < 0$ が成り立つ場合も，目標 $\mathrm{sgn} \circ f$ と視点 $\mathrm{sgn} \circ g$ は ρ に基づく入力 x についてほぼ逆の出力であり，互いに逆相関していると考えられる．上記の 2 つの状況はどちらも目標仮説 f と視点仮説 g が相関しているため，これらを抑制できれば目標仮説 f は視点仮説 g について中立である．そこで，中立性の評価量を以下のように定義する．

定義 1 (+1/-1 汎化中立性リスク)．$f \in \mathcal{F}$ と $g \in \mathcal{G}$ をそれぞれ目標仮説，視点仮説とし，ρ を $(Z, \mathcal{Z})$ 上の確率測度とする．このとき，目標仮説 f の視点仮説 g に関する確率測度 ρ についての+1/-1 汎化中立性リスクは，以下のように定義される．

$$C_{\mathrm{sgn}}(f, g) = \left| \int \mathrm{sgn}(f(x)g(x))d\rho \right|$$

確率測度 ρ が得られないとき，+1/-1 汎化中立性リスク $C_{\mathrm{sgn}}(f, g)$ は事例集合 D_n について経験的に評価される．

定義 2 (+1/-1 経験中立性リスク)．$D_n = \{(x_i, y_i)\}_{i=1}^n \in Z^n$ を与えられた事例集合，$f \in \mathcal{F}$ と $g \in \mathcal{G}$ をそれぞれ目標仮説，視点仮説とする．このとき，目標仮説 f の視点仮説 g に関する事例集合 D_n についての+1/-1 経験中立性リスクは，以下のように定義される．

$$C_{n,\mathrm{sgn}}(f, g) = \frac{1}{n} \left| \sum_{i=1}^n \mathrm{sgn}(f(x_i)g(x_i)) \right| \tag{1}$$

2.2 中立化経験損失最小化

前説で定義した中立性リスクを用いた，教師あり学習において中立化をおこなう枠組みである中立化経験損失最小化 (Neutralized Empirical Risk Minimization : NERM) を提案する．NERM は，経験損失と経験+1/-1 中立性リスクを最小化する．形式的には，以下のように定義される．

$$\min_{f \in \mathcal{F}} \ R_n(f) + \Omega(f) + \eta C_{n,\mathrm{sgn}}(f, g) \tag{2}$$

ここで，$\eta > 0$ は中立化パラメータであり経験損失と中立性リスクの間のトレードオフを制御する．

2.3 +1/-1 中立性リスクの凸緩和

式 (2) の最小化問題は，式 (1) が非凸であるため効率的に解くことができない．そこで，$C_{n,\mathrm{sgn}}(f, g)$ における絶対値を max 関数を用いて緩和し，符号関数を凸な別の関数で緩和することによって，凸緩和した+1/1 中立性リスクを導く．

I を指示関数としたとき，+1/-1 汎化中立性リスクは以下のように 2 つの項に分解できる．

$$\begin{aligned} C_{\mathrm{sgn}}(f, g) = & \Big| \underbrace{\int I(\mathrm{sgn}g(x) = \mathrm{sgn}f(x))d\rho}_{f \text{ と } g \text{ の出力が同じ割合}} \\ & - \underbrace{\int I(\mathrm{sgn}g(x) \neq \mathrm{sgn}f(x))d\rho}_{f \text{ と } g \text{ の出力が違う割合}} \Big| \\ & \coloneqq |C^+_{\mathrm{sgn}}(f, g) - C^-_{\mathrm{sgn}}(f, g)| \end{aligned} \tag{3}$$

+1/-1 汎化中立性リスク $C_{\mathrm{sgn}}(f, g)$ の上界は，$C^+_{\mathrm{sgn}}(f, g)$ と $C^-_{\mathrm{sgn}}(f, g)$ が近ければタイトである．ここから，以下のことが言える．

命題 1. $C^+_{\mathrm{sgn}}(f, g)$，$C^-_{\mathrm{sgn}}(f, g)$ を，式 (3) で定義する．任意の $\eta \in [0.5, 1]$ について，もし

$$C^{\max}_{\mathrm{sgn}}(f, g) \coloneqq \max(C^+_{\mathrm{sgn}}(f, g), C^-_{\mathrm{sgn}}(f, g)) \leq \eta$$

ならば

$$C_{\mathrm{sgn}}(f, g) = |C^+_{\mathrm{sgn}}(f, g) - C^-_{\mathrm{sgn}}(f, g)| \leq 2\eta - 1$$

命題 1 は，$C_{\mathrm{sgn}}(f, g)$ の代わりに $C^{\max}_{\mathrm{sgn}}(f, g)$ を汎化中立性リスクとして用いることができることを示している．次に，$C^{\pm}_{\mathrm{sgn}}(f, g)$ における指示関数を緩和する．

定義 3 (凸緩和汎化中立性リスク)．$f \in \mathcal{F}$ と $g \in \mathcal{G}$ をそれぞれ目標仮説，視点仮説とし，ρ を $(Z, \mathcal{Z})$ 上の確率測度とする．$\psi : \mathbb{R} \to \mathbb{R}^+$ を凸関数とし，

$$C^{\pm}_{\psi}(f, g) = \int \psi(\pm g(x)f(x))d\rho.$$

としたとき，目標仮説 f の視点仮説 g に関する確率測度 ρ についての凸緩和汎化中立性リスクは以下のように定義される．

$$C_{\psi}(f, g) = \max(C^+_{\psi}(f, g), C^-_{\psi}(f, g))$$

凸緩和経験中立性リスクも以下のように定義される．

定義 4 (凸緩和経験中立性リスク)．$D_n = \{(x_i, y_i)\}_{i=1}^n \in Z^n$ を与えられた事例集合，$f \in \mathcal{F}$ と $g \in \mathcal{G}$ をそれぞれ目標仮説，視点仮説とする．$\psi : \mathbb{R} \to \mathbb{R}^+$ を凸関数とし，

$$C^{\pm}_{n,\psi}(f, g) = \frac{1}{n} \sum_{i=1}^n \psi(\pm g(x_i)f(x_i)).$$

としたとき，目標仮説 f の視点仮説 g に関する事例集合 D_n についての凸緩和経験中立性リスクは以下のように定義される．

$$C_{n,\psi}(f,g) = \max(C^+_{n,\psi}(f,g), C^-_{n,\psi}(f,g)).$$

$C^{\pm}_{n,\psi}(f,g)$ は，凸関数 ψ の和であるため凸である．f_1 と f_2 が凸であるならば $\max(f_1(x), f_2(x))$ も凸であることを利用すると，$C_{n,\psi}(f,g)$ は凸であることがわかる．

2.4 凸緩和経験中立性リスクによる NERM

凸緩和経験中立性リスクを用いて，NERM の凸な目的関数は以下のように定式化される．

$$\min_{f\in\mathcal{F}} \; R_n(f) + \Omega(f) + \eta C_{n,\psi}(f,g).$$

経験損失と正則化関数が凸ならば，これは凸最適化問題となる．

3. 汎化中立性リスクバウンド

本章では，NERM の汎化中立性リスクに関する理論的な解析を行う．はじめに，任意の $f \in \mathcal{F}$ における汎化中立性リスクの確率的一様バウンドを導く．次に，NERM の最適な仮説における汎化中立性リスクのバウンドを導く．

3.1 汎化中立性リスクの一様バウンド

Rademacher Complexity は仮説集合 $\mathcal{F}$ の複雑さをはかる指標であり，仮説集合 $\mathcal{F}$ の Rademacher Complexity は以下のように定義される．

$$\mathcal{R}_n(\mathcal{F}) = \mathrm{E}_{D_n,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right]$$

関数 $g : X \to \mathbb{R}$ について，$g\mathcal{F} = \{h : f \in \mathcal{F}, h(x) = g(x)f(x)\ \forall x \in X\}$ とすると，任意の $f \in \mathcal{F}$ における $C_\psi(f,g)$ の確率的一様バウンドは，Rademacher Complexity を用いて以下のように導ける．

定理 1. $C_\psi(f,g)$ と $C_{n,\psi}(f,g)$ を，それぞれ，$g \in \mathcal{G}$ に関する $f \in \mathcal{F}$ の凸緩和汎化中立性リスク，凸緩和経験中立性リスクとし，$\psi : \mathbb{R} \to [0,c]$ をリプシッツ定数 L_ψ のリプシッツ連続な関数であるとする．このとき，少なくとも確率 $1-\delta$ で，すべての仮説 $f \in \mathcal{F}$ について以下を満たす．

$$C_\psi(f,g) \le C_{n,\psi}(f,g) + 2L_\psi\mathcal{R}_n(g\mathcal{F}) + c\sqrt{\frac{\ln(2/\delta)}{2n}}$$

定理 1 により，汎化中立性誤差 $C_\psi(f,g) - C_{n,\psi}(f,g)$ は，n を事例数，δ を信頼パラメータとしたとき，仮説クラス $g\mathcal{F}$ の Rademacher Complexity と $O(\sqrt{\ln(1/\delta)/n})$ によって一様バンドされることを示した．

3.2 NERM の最適な仮説における汎化中立性リスクバウンド

$\hat{f} \in \mathcal{F}$ を NERM の最適な仮説としたとき，以下の条件の基で $\hat{f}$ の経験中立性リスクと汎化中立性リスクのバウンドを導く．

1. 仮説クラス $\mathcal{F}$ に f_0 ($f_0(x) = 0\ \forall x$) が含まれる
2. f_0 の正則化項は $\Omega(f_0) = 0$ (A)

上記の条件は，比較的無理のない条件である．例えば，線形仮説 $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$ において，$W \subseteq \mathbb{R}^D$ を線形仮説クラス，正則化関数を $\Omega(f) = \|\boldsymbol{w}\|_2^2$（$\ell_2^2$ ノルム）としたとき，$\mathbf{0} \in W$ ならば (A) は成り立つ．仮説クラス $\mathcal{F}$ が (A) を満たすとき，$\hat{f}$ の汎化中立性リスクバウンドに関する以下の定理が導ける．

定理 2. $\hat{f}$ は，視点仮説 $g \in \mathcal{G}$，中立化パラメータ η の *NERM* において最適な目標仮説であり，$\psi : \mathbb{R} \to [0,c]$ をリプシッツ定数 L_ψ のリプシッツ関数であるとする．条件 *(A)* が成り立つならば，少なくとも確率 $1-\delta$ で，

$$C_\psi(\hat{f},g) \le \psi(0) + \phi(0)\frac{1}{\eta} + 2L_\psi\mathcal{R}_n(g\mathcal{F}) + c\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

定理 2 の証明のために，以下の系によって $\hat{f}$ の経験中立性リスクの上限を求める．

系 1. 条件 *(A)* が成り立つならば，$\hat{f}$ の凸緩和経験中立性リスクは以下の式でバウンドされる．

$$C_{n,\psi}(\hat{f},g) \le \psi(0) + \phi(0)\frac{1}{\eta}$$

定理 2 は，定理 1 と系 1 を用いることで証明できる．

4. 中立 SVM

4.1 主問題

サポートベクタマシーン (Support Vector Machines : SVMs) [Vapnik 98] は，マージンを基にした 2 値分類の教師あり学習の学習手法である．ソフトマージン SVM は，目標仮説として線形仮説 $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$，損失関数としてヒンジロス $\ell(y, f(x)) = \phi(yf(x)) = [1 - yf(x)]_+$ ($[\cdot]_+ = \max(0,\cdot)$)，正則化項として ℓ_2^2 ノルムを用いる分類器と解釈できる．中立 SVM は NERM における SVM であり，損失関数と正則化項はソフトマージン SVM と同じ，代替関数 ψ はヒンジロスと同じ $\psi(\pm g(x)f(x)) = [1 \mp g(x)f(x)]_+$ を用いる．視点仮説は，任意の仮説を用いることができる．形式的に中立 SVM は NERM 原理に従い，以下のように定式化できる．

$$\min_{\boldsymbol{w},b} \; \sum_{i=1}^{n} 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)]_+ + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 + \eta C_{n,\psi}(\boldsymbol{w},b,g) \quad (4)$$

$$C_{n,\psi}(\boldsymbol{w},b,g) = \max(C^+_{n,\psi}(\boldsymbol{w},b,g), C^-_{n,\psi}(\boldsymbol{w},b,g))$$

$$C^{\pm}_{n,\psi}(\boldsymbol{w},b,g) = \sum_{i=1}^{n}[1 \mp g(\boldsymbol{x}_i)(\boldsymbol{w}^T\boldsymbol{x}_i + b)]_+$$

中立 SVM は損失，正則化，中立化に関するすべての項は凸であるため，目的関数は凸である．中立 SVM の主問題は，式 (4) に subgradient method [Shor 85] を用いることによって解くことができる．

4.2 双対問題とカーネル化

式 (4) の双対問題を導くことで，中立 SVM のカーネル版であるカーネル中立 SVN を求める．スラック変数 $\boldsymbol{\xi}, \boldsymbol{\xi}^{\pm}, \zeta$ により，式 (4) は以下のように表すことができる．

$$\min_{\substack{\boldsymbol{w},b,\\ \boldsymbol{\xi},\boldsymbol{\xi}^{\pm},\zeta}} \; \sum_{i=1}^{n}\xi_i + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 + \eta\zeta \quad (5)$$

$$\text{sub to } \sum_{i=1}^{n}\xi_i^+ \le \zeta, \sum_{i=1}^{n}\xi_i^- \le \zeta, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \le \xi_i,$$

$$1 - v_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \le \xi_i^+, 1 + v_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \le \xi_i^-,$$

$$\xi_i \ge 0, \xi_i^+ \ge 0, \xi_i^- \ge 0, \zeta \ge 0$$

式 (5) のラグランジュ緩和によって，以下のように双対問題を導くことができる．

アルゴリズム 1: SMO-like な中立 SVM の最適化

```
Find γ^1 as the initial feasible solution. Set k = 1
repeat
    Select Working Set B = {i, j} ⊂ {1, ..., 3n} (i ≠ j)
    Update γ^k to γ^{k+1}
    k ← k + 1
until Convergence
```

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}^{\pm}} \lambda\sum_{i=1}^{n} b_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_i x_i^T x_j \tag{6}$$

$$\text{sub to } \sum_{i=1}^{n} a_i = 0, 0 \le \alpha_i \le 1, 0 \le \beta_i^{\pm}, \beta_i^{+} + \beta_i^{-} \le \eta$$

ここで，$b_i = \alpha_i + \beta_i^{+} + \beta_i^{-}, a_i = \alpha_i y_i + \beta_i^{+} v_i - \beta_i^{-} v_i$ である．双対問題において，$\boldsymbol{x}_i^T \boldsymbol{x}_j = k(x_i, x_j)$ とすることによって自然に中立 SVM をカーネル化することができる．

4.3 カーネル中立 SVM の最適化

式 (6) の最適化問題は，*2 次計画問題* (Quadratic Programming : QP) の一種であり，QP のソルバーを用いることによって解くことができるが，メモリの制限により大量のデータに対してスケールしない．そこで，SVM の最適化手法としてよく知られる *Sequencial Minimal Optimization (SMO)* を，中立 SVM に適用する．SMO ではメモリの消費を抑えるために，1 回の更新においてワーキングセットとよばれるパラメータの部分集合のみを変更する．中立 SVM を解くための SMO-like なアルゴリズムをアルゴリズム 1 に示す．アルゴリズムにおいて，$\boldsymbol{\gamma} = (\alpha_1, ..., \alpha_n, \beta_1^{+}, ..., \beta_n^{+}, \beta_1^{-}, ..., \beta_n^{-})^T$ である．紙数の都合上，各ステップの詳細は省略する．

5. 実験

UCI Repository[Bache 13] の German Credit データセットにおいて，CV2NB，PR，ηLR と中立 SVM の比較を行った．German Gredit は 20 属性をもつ 1000 事例からなり，目標 y は属性 credit risk，視点 v は属性 foreign worker を用いた．正則化パラメータ，カーネル関数やそのパラメータなどは，各アルゴリズムにおいて中立化をしない場合に 5 分割交差検定で一番精度が良いものを選択した．PR，ηLR，中立 SVM における中立化パラメータは，PR は $\{0, 0.01, 0.05, ..., 100\}$，$\eta$LR は $\{0, 5\times 10^{-5}, 1\times 10^{-4}, ..., 0.5\}$，中立 SVM は $\{0, 0.01, 0.05, ..., 100\}$ を用いた．分類の評価は AUC，中立性の評価は経験中立性リスク $C_{n,\mathrm{sgn}}(f, g)$ を用いた．評価量は，5 分割交差検定を別々の 10 の分割における平均を算出した．

結果 図 1 に，個々の中立化パラメータにおける実験結果を示す．図において横軸が AUC，縦軸が $C_{n,\mathrm{sgn}}(f, g)$ を表しており，右下に行くほど良い結果である．各アルゴリズムにおいて，ほかの点よりも ACU，$C_{n,\mathrm{sgn}}(f, g)$ どちらも低くなる点は削除している．CV2NB，PR，ηLR の点よりも，中立 SVM の線が右下にあることがわかる．従って，提案法が AUC と $C_{n,\mathrm{sgn}}(f, g)$ のよいトレードオフを実現しているといえる．

6. まとめ

本稿では，2 値の分類を行う ERM について，与えられた 2 値の視点に対する中立化を行う枠組みである NERM を提案した．NERM について，以下の 3 つのことを示した．(1) 中立化を行う枠組みとして NERM を定義し，凸緩和をすることによって凸計画問題として定式化できることを示した．既存手法は非凸で局所解しか得られなかったことに対し，NERM は大域的最適解が保証される．(2)NERM の枠組みにおいて，汎化中立性に関する理論的解析を行った．理論解析では，中立性の汎化近似誤差が target hypothesis の仮説クラス $\mathcal{F}$ の Rademacher Complexity と $O_p(1/\sqrt{n})$ でバウンドできることを示した．また，中立化パラメータ η と汎化中立性のバウンドの関係を導出した．(3)NERM に乗っ取ったアルゴリズムとして，中立 SVM を紹介した．中立 SVM は，双対問題を求めることでカーネル化が可能であることを導出した．

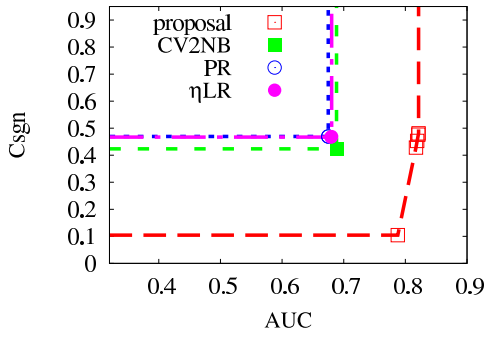

図 1: CV2NB，PR，ηLR，中立 SVM (proposal) の比較実験の結果．縦軸は AUC，横軸は $C_{n,\mathrm{sgn}}(f, g)$ を示している．

謝辞

本研究は，JST CREST「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」領域におけるプロジェクト「自己情報コントロール機構を持つプライバシ保護データ収集・解析基盤の構築と個別化医療・ゲノム疫学への展開」の助成を受けました．

参考文献

[Bache 13] Bache, K. and Lichman, M.: UCI Machine Learning Repository (2013)

[Calders 10] Calders, T. and Verwer, S.: Three Naive Bayes Approaches for Discrimination-Free Classification, *Data Mining and Knowledge Discovery*, Vol. 21, No. 2, pp. 277–292 (2010)

[Fukuchi 13] Fukuchi, K., Sakuma, J., and Kamishima, T.: Prediction with Model-Based Neutrality, in Blockeel, H., Kersting, K., Nijssen, S., and Zelezný, F. eds., *ECML/PKDD (2)*, Vol. 8189 of *Lecture Notes in Computer Science*, pp. 499–514, Springer (2013)

[Kamishima 12] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: Fairness-aware Classifier with Prejudice Remover Regularizer, in *in Proceedings of the ECML/PKDD2012, Part II*, Vol. LNCS 7524, pp. 35–50, Springer (2012)

[Pariser 11] Pariser, E.: *The Filter Bubble: What The Internet Is Hiding From You*, Viking, London (2011)

[Shor 85] Shor, N. Z., Kiwiel, K. C., and Ruszcayǹski, A.: *Minimization Methods for Non-differentiable Functions*, Springer-Verlag New York, Inc., New York, NY, USA (1985)

[Vapnik 98] Vapnik, V. N.: Statistical learning theory (1998)

[Zemel 13] Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C.: Learning Fair Representations, in *ICML (3)*, Vol. 28 of *JMLR Proceedings*, pp. 325–333, JMLR.org (2013)