2B5-OS-15b-2

匿名化データからのロバスト分類学習とその汎化誤差解析

Generalization Error Analysis of Robust Classification Learning with Anonymized Data

小林星平[*1] 佐久間淳[*1*2]
Shohei Kobayashi 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]科学技術振興機構 CREST
Japan Science and Technology Agency, CREST

We regard the data manipulation process caused by privacy preservation as a perturbation in robust optimization. Then, we apply this idea for the analysis of *k-anonymity*, which is recognized as a fundamental privacy model. We specifically investigate *classification*. Using previous results of statistical learning theory for predictors trained by robust optimization, we proved that upper bound of expected loss of classification is robust to the perturbation introduced for the purpose of privacy preservation. Furthermore, we provide a theoretical interpretation of the relation between a data manipulation for privacy preservation and utility of data analysis. According to our result, the privacy and utility are not necessarily in a trade-off relationship. In the experiments, we empirically evaluate the relation of anonymity and utility with several realworld datasets.

1. はじめに

技術やサービスの発展に伴い, 大量のデータが蓄積され利用されるようになった. 例えば, 病院では患者の疾患, 治療の記録が保存され, ショッピングサイトでは購買履歴や商品の閲覧履歴などを収集している. これらのデータを解析することで, 将来の疾患リスクの予測や商品の推薦など, 多くの有用なサービスへの利用が可能となる. また, これらのデータやサービスを繋ぎ合わせ連携させることによって得られる価値は大きい. しかし, 病院やサービス提供者が持つデータを解析機関に提供したり, 連携のために別のサービス提供者へデータを受け渡す場合, 疾患や治療の記録, 住所や購買履歴など他人に知られたくない情報が含まれるために, これらのデータが一体誰のものなのかを一意に特定できないようにする必要がある. そのための技術がデータ匿名化である.

匿名化では, 定義した指標に基づき元のオリジナルデータに対して個人が識別できないようにデータを加工する. データの有用性を元データからの歪みと解釈するならば, この加工によって有用性は一般に低下する. この場合, 匿名性と有用性の間にはトレードオフがあると考えられ, 匿名性と有用性を両立させることは一般に困難である. このことについて, [Sankar 13] では元データからの歪みを有用性と解釈した場合の, 有用性と匿名性のトレードオフについて情報理論的アプローチから理論的に解析を行っている. 一方で, 匿名化のためのデータの加工がデータ解析の出力に与える変化量を有用性と解釈した場合, この匿名性と有用性の間のトレードオフは, 実験的な評価はされてきたが, 理論的解析はほとんどされていない. 例えば, [Mohammed 11] は分類で用いるようなデータセットについて, データセットに摂動を加えることでその分類結果が差分プライバシを満たすような手法を提案している. このときの分類精度, つまり有用性と匿名性の関係は実験での評価のみとなっており, 理論的に解析されていない.

連絡先: 小林星平, 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, kobayashi@mdl.cs.tsukuba.ac.jp

既存研究では匿名性と有用性の関係は主に実験的評価がされてきたが, 実験的評価ではデータやデータ解析手法に依存する. 匿名化がデータ解析に与える影響をより一般的に議論するためには, 匿名性と有用性についての理論的解析が必要である.

1.1 貢献

我々は, 匿名化データを用いた協調フィルタリングに関する研究において, 有用性と匿名性が単純なトレードオフの関係ではないことを実験的に示した [納 13]. また, 線形回帰において, データ解析の精度を有用性とした場合の匿名性と有用性の関係を理論的に示した [小林 14]. 本稿では Support Vector Machine (SVM) での分類学習に注目し, 匿名性と有用性の関係を統一的に記述する. さらに, 線形回帰の場合と同様に, この関係が理論的にも単純なトレードオフの関係にないことを示す. 本稿の貢献を以下にまとめる.

- 匿名化のためのデータ変更量をロバスト最適化における摂動と解釈し, これに対しロバスト最適化の枠組みで匿名化データから学習する方法を導入した.
- 有用性と匿名性をカバリングナンバーにより統一的に記述し, 関係付けた. さらに, ロバスト分類学習について, この有用性と匿名性の関係が必ずしもトレードオフとならないことをカバリングナンバーによる記述に基づき, 理論的に示した.

2. 準備

2.1 経験損失最小化

経験損失最小化とは, 統計的学習理論における予測モデルの構築において標準的に用いられる定式化である. $\mathcal{Z}$ をデータ空間とする. 各データ $z \in \mathcal{Z}$ は入力と目標値のペア $z = (\boldsymbol{x}, y)$ で表される. 観測された, 学習に用いる n 個のデータ s_i の集合を訓練集合と呼び, $\boldsymbol{s} = \{s_1, ..., s_n\} \in \mathcal{Z}^n$ で表す. また, 訓練データ s_i の $\boldsymbol{x}$ 要素を $s_{i|\boldsymbol{x}}$, y 要素を $s_{i|y}$ で表記する. 観測されているが訓練に用いなかったデータをテストデータと呼び, その集合を $\boldsymbol{t}$ で表す. 入力から目標値を予測する関数を仮説 h

と呼ぶ. この仮説の集合を仮説空間 $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$ で表す. 学習とは, 訓練集合を用いて以下に示す意味で最適な仮説を求めることに他ならず, 学習アルゴリズム $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$ は訓練集合からこの仮説空間への写像として記述できる. 訓練集合 $\boldsymbol{s}$ で学習したアルゴリズムを $\mathcal{A}_{\boldsymbol{s}} \in \mathcal{H}$ と書く. 各仮説とデータ点の間に損失 $l(\cdot,\cdot) : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ を与える. この損失は非負で, ある正の実数値 $M \in \mathbb{R}$ が上限であると仮定する. データが未知の分布 μ から独立同分布 (independent identically ditributed, 以下 i.i.d.) で生成されるとする. このとき, 期待損失 $\hat{l}(\cdot)$ は $\hat{l}(\mathcal{A}_{\boldsymbol{s}}) \triangleq \mathbb{E}_{z\sim\mu} l(\mathcal{A}_{\boldsymbol{s}}, z)$ と定義される. 期待損失は未知のデータに対する予測精度と解釈でき, これを改善するためには期待損失を最小化することが望ましいが, 一般にデータの分布 μ は未知であるため, 期待損失を直接評価することはできない. そこで, 観測されたデータに対する損失を最小化する. これを経験損失と呼ぶ. 経験損失は訓練集合 $\boldsymbol{s}$ と, $\boldsymbol{s}$ で学習した仮説 $\mathcal{A}_{\boldsymbol{s}}$ によって $l_{\mathrm{emp}}(\mathcal{A}_{\boldsymbol{s}}) \triangleq \frac{1}{n}\sum_{s_i\in\boldsymbol{s}} l(\mathcal{A}_{\boldsymbol{s}}, s_i)$ と定義される. 与えられた入力と目標値の関係を最も良く表現する仮説として, この経験損失を最小化する仮説を求める学習の枠組みを経験損失最小化と呼ぶ.

SVM では, パラメータ $\boldsymbol{w}$ に対し, 予測は $y = \langle \boldsymbol{w}, s_{|\boldsymbol{x}}\rangle + d$ で与える. また, 損失関数を $l(\boldsymbol{w}, z) = [1 - z_{|y}(\langle \boldsymbol{w}, z_{|\boldsymbol{x}}\rangle + d)]_+$ とする. ここで, $\langle\cdot,\cdot\rangle$ は内積, $[\cdot]_+$ はヒンジ損失を表す. 経験損失最小化の枠組みでは, これを最小とするような $\boldsymbol{w}$, つまり

$$\min_{\boldsymbol{w},d}\left\{\frac{1}{n}\sum_{i=1}^{n}[1 - s_{i|y}(\langle \boldsymbol{w}, s_{i|\boldsymbol{x}}\rangle + d)]_+\right\}$$

の解が SVM において入力と目標値の関係を最も良く表現する仮説であると考える.

2.2 カバリングナンバー

カバリングナンバーは統計的学習理論で導入された, 仮説集合の複雑さ, 容量を表す概念である. また, カバリングナンバーはデータ集合の複雑さの定量化にも用いられる [Xu 10]. 本稿でも, カバリングナンバーをデータ集合の複雑さを表現することに利用する. ϵ-カバリングナンバー $\mathcal{N}(\epsilon, T, \rho)$ とは, 距離 ρ における半径 ϵ の超球で空間 T を覆うのに必要な最小の超球の個数のことである. 形式的には以下で定義される.

定義 1 (ϵ-カバリングナンバー [Vaart 00])**.** 距離空間 S における, 距離関数を ρ とする. また, $T \subset S$ を S の部分集合とする. $\forall t \in T$ において, $\rho(t, \hat{t}) \le \epsilon$ なる $\hat{t} \in \hat{T}$ が存在するとき, $\hat{T}$ は T の ϵ-カバーと呼ぶ. このとき, T の ϵ-カバリングナンバーは次のように定義される.

$$\mathcal{N}(\epsilon, T, \rho) = \min\{|\hat{T}|\}.$$

2.3 プライバシモデル

データを解析機関に提供したり, 連携のために別のサービス提供者へデータを受け渡す場合, 疾患や治療の記録, 住所や購買履歴など, 他人に知られたくない情報が含まれることから, これらのデータが一体誰のものなのかを一意に特定できないようにするためデータ匿名化が必要となる. 匿名化では, 定義した指標に基づき元のオリジナルデータに対して個人が識別できないようにデータを加工する. この指標の一つとして, データセット中で k 人と見分けがつかないようにデータを加工する k-匿名性 [Sweeney 02] がある.

k-匿名性は任意のデータについて定義された一般的な概念であるが, 本稿では扱うデータを実数値のベクトルに限定し, 匿名性の定義もこれに基づいて行う. 本稿では, 与えられたデータに変化を加えるモデル (入力プライバシ) を扱う.

定義 2 (k 匿名性)**.** $\boldsymbol{s}$ の任意のベクトルにおいて, 同じベクトルが $\boldsymbol{s}$ の中に少なくとも他に $k-1$ 個あるとき, $\boldsymbol{s}$ は k-匿名性を持つ.

つまり, 訓練集合中に同じベクトルが少なくとも k 個ずつあるとき, この訓練集合は k-匿名性を満たしている.

2.4 カバリングナンバーによる 匿名性の表現

カバリングナンバーはデータ集合の複雑さの定量化にも利用することができる. 一定のプライバシ保護 (匿名性) を実現するために必要なデータの変更量を, カバリングナンバーを用いて表現する. k-匿名性の場合は, 次のように表現される.

命題 1 (カバリングナンバーによる k 匿名性表現)**.** $\mathcal{Z}$ における距離関数を ρ とする. $\boldsymbol{s} \in \mathcal{Z}^n$ について $\{\hat{s}_1, ..., \hat{s}_{\mathcal{N}(\epsilon_k, \boldsymbol{s}, \rho)}\}$ は $\boldsymbol{s}$ の ϵ_k-カバーである. $C_i = \{z \in \mathcal{Z} \mid \rho(z, \hat{s}_i) \le \epsilon_k\}$ において, 各 C_i は互いに素ですべての i で $|C_i| \ge k$ とする. このとき, C_i の各要素を $\hat{s}_i$ で置き換えれば $\boldsymbol{s}$ は k-匿名性を持つ.

この証明はカバリングナンバーの定義より自明である. 例として, n 個の実数値ベクトル $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n \mid \boldsymbol{x}_i \in \mathbb{R}^d\}$ について k-匿名化することを考える. 距離空間として $\mathbb{R}^d$ を考え, その要素の集合 T が n 個の実数値ベクトルであるとする. また, 距離としてユークリッド距離を用いる. 命題 1 より, $\mathcal{N}(\epsilon_k, T, \|\cdot\|_2)$ 個の超球により k-匿名性が達成される. このとき, 各データ点の自身を含むカバーまでの距離は ϵ_k 以下となる.

3. ロバストな学習アルゴリズムの汎化性能

前章ではカバリングナンバーによってプライバシモデルを表現した. 本章では, algorithmic robustness と呼ばれる, 学習アルゴリズムが生成する仮説の頑健さに関する性質を通じて, 学習アルゴリズムの期待損失がカバリングナンバーと関係付けられることを示す. 有用性として期待損失を用いたとき, このことから, プライバシと有用性の関係がカバリングナンバーを通じて記述できることを 4. 章で示す.

学習アルゴリズムの性質から期待損失の上界を求める手法が多く研究されている. 本稿では学習アルゴリズムの期待損失のバウンドを求める手法として, Xu と Mannor によって提案された学習アルゴリズムのロバスト性 (algorithmic robustness) を基にしたアプローチ [Xu 10] を用いる.

3.1 Algorithmic robustness

Algorithmic robustness とは, 直感的には, テスト集合が訓練集合と「似ている」とき, 訓練集合で学習した学習アルゴリズムによるテスト誤差が訓練誤差に近ければ「学習アルゴリズムはロバストである」という学習アルゴリズムの性質である. 学習アルゴリズムのロバスト性は, 訓練データに含まれる (小さな) 摂動への鈍感さという性質を意味する.

定義 3 (Algorithmic robustness [Xu 10])**.** $\mathcal{Z}$ が K 個の互いに素な有限の部分集合 $\{C_i\}_{i=1}^K$ に分割されているとき, すべての $s \in \boldsymbol{s}$ で

$$s, z \in C_i \Rightarrow |l(\mathcal{A}_{\boldsymbol{s}}, s) - l(\mathcal{A}_{\boldsymbol{s}}, z)| \le \epsilon(\boldsymbol{s})$$

が成り立つならば, 学習アルゴリズム $\mathcal{A}$ は $(K, \epsilon(\boldsymbol{s}))$-*algorithmic robust* である.

3.2 カバリングナンバーによる Algorithmic robustness の表現

Algorithmic robustness は前述の匿名性と同様に, 以下のようにカバリングナンバーによって記述される.

定理 1 ([Xu 10]). $\gamma > 0$ かつ ρ を $\mathcal{Z}$ の距離関数とする. $\mathcal{N}(\gamma/2, \mathcal{Z}, \rho) < \infty$ であり, 学習アルゴリズム $\mathcal{A}$ が次の式を満たすならば

$$|l(\mathcal{A}_{\boldsymbol{s}}, z_1) - l(\mathcal{A}_{\boldsymbol{s}}, z_2)| \leq e(\boldsymbol{s}), \forall z_1, z_2 : z_1 \in \boldsymbol{s}, \rho(z_1, z_2) \leq \gamma$$

学習アルゴリズム $\mathcal{A}$ は $(\mathcal{N}(\gamma/2, \mathcal{Z}, \rho), e(\boldsymbol{s}))$-*algorithmic robust* である.

3.3 Algorithmic robustness に基づく汎化誤差解析

標準的な学習の設定, 即ち訓練集合 $\boldsymbol{s}$ が未知の分布 μ から i.i.d. で生成される n 個のサンプルで構成されていることを考える. 学習アルゴリズム $\mathcal{A}$ が $(K, \epsilon(\boldsymbol{s}))$-algorithmic robust を持つとき, $\mathcal{A}$ の期待損失のバウンドは定理 2 で示される.

定理 2 ([Xu 10]). $\boldsymbol{s}$ が n 個の *i.i.d.* によって生成されたサンプルで構成されており, アルゴリズム $\mathcal{A}$ が $(K, \epsilon(\boldsymbol{s}))$-*algorithmic robust* であるとき, すべての $\delta > 0$ において, 少なくとも $1 - \delta$ の確率で以下が成り立つ.

$$\hat{l}(\mathcal{A}_{\boldsymbol{s}}) \leq l_{emp}(\mathcal{A}_{\boldsymbol{s}}) + \epsilon(\boldsymbol{s}) + M\sqrt{\frac{2K\ln 2 + 2\ln(1/\delta)}{n}}. \quad (1)$$

(1) 式から, 訓練集合のサイズ n を大きくすれば期待損失が小さくできることが分かる. また, algorithmic robustness の $\epsilon(\boldsymbol{s}), K$ が増えることは期待損失を大きくしてしまうために, より小さい $\epsilon(\boldsymbol{s}), K$ を持つ学習アルゴリズムが高い有用性を持つことが分かる.

4. 匿名データからの学習とその汎化誤差解析

データにノイズなどによる変化, 摂動が加わることを考える. 事例 s_i の $\boldsymbol{x}$ 成分に加わる摂動を $\boldsymbol{u}_i$ とする. 経験損失最小化を考えた場合は, 観測された特定の摂動 $\{\boldsymbol{u}_i\}_{i=1}^n$ に対して経験損失を最小化する.

$$\min_{\boldsymbol{w},d}\left\{\frac{1}{n}\sum_{i=1}^{n}[1 - s_{i|y}(\langle \boldsymbol{w}, s_{i|\boldsymbol{x}} - \boldsymbol{u}_i\rangle + d)]_+\right\}.$$

ここで匿名性を満たすためにデータに加える変化を摂動であると解釈すれば, $\{s_{i|\boldsymbol{x}} - \boldsymbol{u}_i\}_{i=1}^n$ は匿名化データに対応する. しかし, 期待損失を有用性と捉えた場合, 特定の摂動に対して経験損失を最小化することは必ずしも期待損失を小さくしない. そこで, 経験損失を最大にする最悪の摂動に対する最良の仮説を考えることで, 仮説が特定の摂動のみに依存しないように学習させることを考える. 最悪の摂動に対して, 最小の経験損失を与える仮説は, ロバスト最適化により求めることができる.

以下では, 匿名化データ $\{s_{i|\boldsymbol{x}} - \boldsymbol{u}_i\}_{i=1}^n$ から直接学習するのではなく, ロバスト最適化を用いて学習したときの期待損失を理論的に評価する.

4.1 ロバスト SVM

訓練データに加わる不確実性を $\mathcal{U} = \{(\boldsymbol{u}_1, ..., \boldsymbol{u}_n) \mid \sum_{i=1}^n \|\boldsymbol{u}_i\|_2 \leq c\}$ とする. このとき, ロバスト SVM は次のように定式化される.

$$\min_{\boldsymbol{w},d}\left\{\max_{(\boldsymbol{u}_1,...,\boldsymbol{u}_n)\in\mathcal{U}}\frac{1}{n}\sum_{i=1}^{n}[1 - s_{i|y}(\langle \boldsymbol{w}, s_{i|\boldsymbol{x}} - \boldsymbol{u}_i\rangle + d)]_+\right\}. \quad (2)$$

これは, 次の式と等価である [Xu 09].

$$\min_{\boldsymbol{w},d}\left\{c\|\boldsymbol{w}\|_2 + \frac{1}{n}\sum_{i=1}^{n}[1 - s_{i|y}(\langle \boldsymbol{w}, s_{i|\boldsymbol{x}}\rangle + d)]_+\right\}. \quad (3)$$

	事例数	特徴数
人工データ (線形分離可能)	400	2
人工データ (線形分離不可能)	400	2
wdbc *1	569	31
pima *1	733	8
ilpd *1	580	10

表 1: データセットの仕様

(3) 式は任意の γ で $(2\mathcal{N}(\gamma/2, \mathcal{X}, \|\cdot\|_2), \gamma/\sqrt{c})$-algorithmic robustness である [Xu 10].

4.2 匿名化されたデータからのロバスト SVM とその期待損失

カバリングナンバーによってその匿名化のためのデータ変化量が表された匿名化データから学習したロバストな SVM について, algorithmic robustness に基づいて汎化誤差解析を行う.

定理 3. 命題 *1* に基づき k 匿名化されたデータ $\boldsymbol{s}'$ から学習したロバスト *SVM* アルゴリズムの期待損失は $1 - \delta$ の確率で

$$\hat{l}(\mathcal{A}_{\boldsymbol{s}'}) \leq l_{emp}(\mathcal{A}_{\boldsymbol{s}'}) + \frac{\gamma}{\sqrt{n\epsilon_k}} + M\sqrt{\frac{2\mathcal{N}(\gamma/2, \mathcal{X}, \|\cdot\|_2)\ln 2 + 2\ln(1/\delta)}{n}}. \quad (4)$$

Proof. 定理 2 より, $(2\mathcal{N}(\gamma/2, \mathcal{X}, \|\cdot\|_2), \gamma/\sqrt{c})$-algorithmic robustness を持つ (2) 式の期待損失のバウンドは

$$\hat{l}(\mathcal{A}_{\boldsymbol{s}'}) \leq \mathrm{lemp}(\mathcal{A}_{\boldsymbol{s}'}) + \frac{\gamma}{c} + M\sqrt{\frac{2\mathcal{N}(\gamma/2, \mathcal{X}, \|\cdot\|_2)\ln 2 + 2\ln(1/\delta)}{n}}.$$

また, 命題 1 より, データ i への k-匿名化のためのデータ操作量 $\boldsymbol{u}_i$ はすべての i で $\|\boldsymbol{u}_i\|_2 \leq \epsilon_k$ を満たす. このことから, データの変化量の総和は $\sum_{i=1}^n \|\boldsymbol{u}_i\|_2 \leq n\epsilon_k$ となる. よって, 摂動の制限 $c = n\epsilon_k$ のとき, k-匿名化された訓練集合 $\boldsymbol{s}'$ で学習したアルゴリズムの期待損失のバウンドは式 (4) となる. □

(4) 式は, k-匿名化したデータで学習した際の期待損失の上限を示している. また, この式は ϵ_k が大きくなり, 匿名化のためにそれぞれのデータの変化量が増えると期待損失が小さくなる場合があることを示唆している. つまり, 匿名化を行うことはデータ解析における有用性を必ずしも劣化させない. しかし, ϵ_k が正則化パラメータと関係していることからも分かるように, 大き過ぎる ϵ_k は過度な汎化をもたらすために経験損失が大きくなり, それにつれて期待損失も大きくなってしまう.

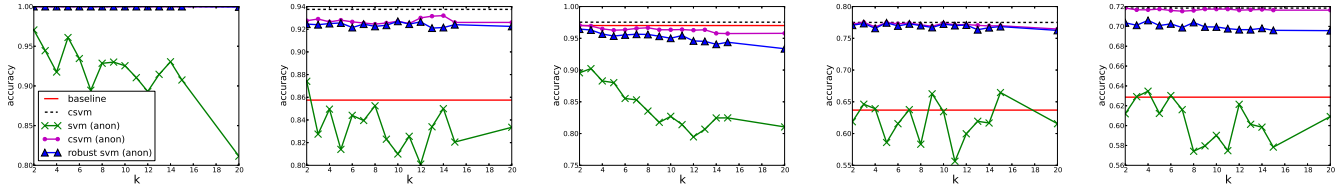
5. 実験

協調フィルタリング, 線形回帰では匿名化したデータで学習した結果, オリジナルデータで学習した場合よりも予測精度が高いことが実験で示されている [納 13, 小林 14]. 分類においても, 同様の結果が得られるか確認するために実験を行った.

5.1 データセットと評価方法

実験用のデータセットとして, 表 1 に示した分類学習のためのデータセットを用いる. ただし, pima と ilpd データセットは欠損値を含む事例を削除している. データセットは前処

*1 http://archive.ics.uci.edu/ml/datasets/



(a) Synthetic (separable) (b) Synthetic (inseparable) (c) wdbc (d) pima (e) ilpd

図 1: 人工データ, wdbc, pima, ilpd データセットに対する実験結果. k を変化させた匿名化データで学習したロバスト SVM の正答率の変化を robust svm (anon) として示す. 比較として, 匿名化されていないオリジナルデータから学習した SVM の正答率を base line, 匿名化したデータで学習した SVM の正答率を svm (anon) として示す. また, 匿名化されていないオリジナルデータから, 正則化パラメータをチューニングし学習した SVM の正答率を csvm, 匿名化したデータで正則化パラメータをチューニングし学習した SVM の正答率を csvm (anon) として示す.

理として各特徴の値が $[-1, 1]$ の範囲となるようにスケーリングを行っている. また, k-匿名化アルゴリズムとして Lin らの OKA-algorithm [Lin 08] を用いる. 予測精度の評価では正答率を精度の指標として用いる. 正答率は, k-匿名化した訓練集合 $\boldsymbol{s}'$ で学習した仮説 $\boldsymbol{w}(\boldsymbol{s}')$ と, テスト集合 $\boldsymbol{t}$ を用いて $accuracy = \frac{1}{|\boldsymbol{t}|}\sum_{i=1}^{|\boldsymbol{t}|} \mathbf{1}(t_{i|y} = (\langle \boldsymbol{w}(\boldsymbol{s}'), t_{i|\boldsymbol{x}} \rangle + d)$ として算出される. ただし, $\mathbf{1}(\cdot)$ は条件が真であるとき 1, それ以外は 0 を返すような関数である. ここで $|\boldsymbol{t}|$ はテスト集合に含まれるデータ数を表している. 10 分割交差検定を行い, 各分割において訓練データの匿名化を乱数の種を変え 5 回ずつ行った結果の平均を取っている. 比較として, 匿名化されていないオリジナルデータから学習した SVM の正答率を base line, 匿名化したデータで学習した SVM の正答率を svm (anon) として示す. また, 匿名化されていないオリジナルデータから, 正則化パラメータをチューニングし学習した SVM の正答率を csvm, 匿名化したデータで正則化パラメータをチューニングし学習した SVM の正答率を csvm (anon) として示す. また, ロバスト SVM では正則化パラメータ c の値に, データ変更量の総和 $\sum_{i=1}^{n} \|\boldsymbol{u}_i\|_2$ を用いている.

5.2 実験結果

各データセットに対しての実験結果を図 1 (a) から (e) にそれぞれ示す. (b), (d), (e) では robust svm は base line よりも良い精度となり, 匿名化データで学習した方がオリジナルデータで学習するよりも良い精度となる場合があることが実験的に示された. それに対し, (c) では匿名性パラメータである k が大きくなるにつれ精度が悪くなっている. これは, オリジナルデータで学習した base line と csvm の精度に大きな差が無いため, (a) と同様, (c) もほぼ線形に分類できるデータであると考えられ, そのために匿名性の増加にあわせて正則化を強くしていくロバスト SVM では, 過度な正則化のために精度が悪くなってしまったのだと考えられる.

6. まとめと今後の課題

本稿では, 有用性とある種のプライバシモデルをカバリングナンバーを用いて統一的に記述した. これにより, 有用性と匿名性について, カバリングナンバーを通じてこれらの関係を記述した. さらに, SVM による分類学習について, 匿名化データからの学習が必ずしも予測精度を劣化させないことを理論的に示した. このことから, 有用性と匿名性の関係が必ずしもトレードオフとならないことを示した. また, SVM において, 有用性と匿名性の関係が単純なトレードオフではなく, 匿名化されていないオリジナルデータでの学習よりも, 匿名化データでの学習の方が有用性が高くなる場合があることを実験により示した. 今後の課題として, k-匿名化以外の差分プライバシなど他のプライバシモデルに対しても理論的, 実験的評価を行うことを考えている.

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

[Lin 08] Lin, J.-L. and Wei, M.-C.: An Efficient Clustering Method for K-anonymization, in *Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society*, PAIS '08, pp. 46–50, New York, NY, USA (2008), ACM

[Mohammed 11] Mohammed, N., Chen, R., Fung, B. C., and Yu, P. S.: Differentially Private Data Release for Data Mining, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pp. 493–501, New York, NY, USA (2011), ACM

[Sankar 13] Sankar, L., Rajagopalan, S. R., and Poor, H. V.: Utility-Privacy Tradeoffs in Databases: An Information-Theoretic Approach, *IEEE Transactions on Information Forensics and Security*, Vol. 8, No. 6, pp. 838–852 (2013)

[Sweeney 02] Sweeney, L.: K-anonymity: A Model for Protecting Privacy, *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, Vol. 10, No. 5, pp. 557–570 (2002)

[Vaart 00] Vaart, van der A. and Wellner, J.: *Weak Convergence and Empirical Processes*, Springer (2000)

[Xu 09] Xu, H., Caramanis, C., and Mannor, S.: Robustness and regularization of support vector machines, *The Journal of Machine Learning Research*, Vol. 10, pp. 1485–1510 (2009)

[Xu 10] Xu, H. and Mannor, S.: Robustness and Generalization, in *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel*, Omnipress (2010)

[小林 14] 小林 星平, 佐久間 淳 : 匿名化データからのロバスト線形回帰とその汎化誤差解析, 暗号と情報セキュリティシンポジウム (2014)

[納 13] 納 竜也, 川本 淳平, 佐久間 淳 : 離散属性付き高次元数値属性のクラスタリングによる匿名化, 暗号と情報セキュリティシンポジウム (2013)