CD-006

離散値属性を持つインスタンスに対する尤度比推定法の人工データによる有効性評価
Synthetic Data-based Evaluation of a Likelihood Ratio Estimator for Instances with Discrete Attributes

菊地 真人[1)]　吉田 光男[2)]　梅村 恭司[3)]　大囿 忠親[1)]
Masato Kikuchi　Mitsuo Yoshida　Kyoji Umemura　Tadachika Ozono

あらまし

離散値属性を持つインスタンスからなるデータセットは，分類や回帰でよく用いられる．尤度比は分類タスクで利用される統計量だが，インスタンスに対する尤度比推定は難しい．なぜなら，属性値の出現頻度に基づく素朴な推定法は，しばしば尤度比を過大に見積もるためである．我々は頻度の低さに応じて尤度比を低めに見積もる “保守的な推定法” を確立した．しかし，この推定法はインスタンスの尤度比推定へ単純に適用できない．そこで，インスタンスの尤度比に対する保守的な推定法の適用方法，および推定に不要な属性の悪影響を抑えるための特徴重み付け法の導入方法を提案する．属性値の頻度と有用性を制御できる人工データを用いた実験により，提案法が効果的な状況を明らかにする．

1　はじめに

複数の属性値を持つインスタンスからなるデータセットは，分類，回帰，推薦といった機械学習の主要タスクで用いられてきた．また，尤度比は分類や統計検定で古くから用いられており，二値分類における最適性が理論証明されるなど重要な統計量である [1]．本稿では，インスタンスに対する尤度比を出現頻度から推定する問題に取り組む．A_k $(k = 1, 2, \ldots, n)$ を属性，$a_k \in A_k$ を頻度の数え上げが可能な属性値（離散値）とする．このとき，インスタンス $y = \langle a_1, a_2, \ldots, a_n \rangle$ に対する尤度比は

$$r(y) = \frac{p_{\mathrm{nu}}(y)}{p_{\mathrm{de}}(y)}$$

と定義される．“de” と “nu” は，尤度比の分母と分子を表す添え字である．y 自体の頻度を得ることは困難なため，$r(y)$ の推定には何らかの工夫が必要になる．

簡単な工夫としてナイーブベイズ分類器のように，a_k の出現が $a_{k'}$ $(k \neq k')$ の出現と統計的独立と仮定し，$r(y)$ を a_k の尤度比 $r(a_k)$ の積

$$r(y) = \prod_{k=1}^{n} r(a_k)$$

として表現する方法がある．これにより，個々の a_k が出現していれば，$r(y)$ の推定ができるようになる．しかしながら，このアプローチには二つの問題がある．第一に，頻度を用いて素朴に $r(a_k)$ を推定すると，$r(a_k)$ の推定値が過大に見積もられる場合がある．$r(y)$ は $r(a_k)$ の積であるから，$r(a_k)$ の過大推定が $r(y)$ の最終的な推定結果に悪影響を及ぼしてしまう．第二に，$r(y)$ の推定に有用ではない属性が混入していた場合，同様に $r(y)$ の推定に悪影響を及ぼす．

1）　名古屋工業大学大学院 情報工学専攻
2）　筑波大学 ビジネスサイエンス系
3）　豊橋技術科学大学 情報・知能工学系

表 1　尤度比の推定例．$\widehat{r}(x)$ の λ は 10^{-5} とした．

x	観測頻度				$r_{\mathrm{MLE}}(x)$	$\widehat{r}(x)$
	n_{de}	$f_{\mathrm{de}}(x)$	n_{nu}	$f_{\mathrm{nu}}(x)$		
x_1	10^7	2,000	10^4	100	50	47.6
x_2	10^7	20	10^4	1	50	8.3
x_3	10^7	20	10^4	2	100	16.7

第一の問題を緩和するため，頻度に応じて尤度比を低めに見積もる推定法 [2] を $r(a_k)$ の推定に用いる．以降，推定値をあえて低めに見積もることを “保守的な推定” と呼称する．まず，頻度による素朴な推定法 $r_{\mathrm{MLE}}(x)$ は

$$r_{\mathrm{MLE}}(x) = \frac{\widehat{p}_{\mathrm{nu}}(x)}{\widehat{p}_{\mathrm{de}}(x)}, \quad \widehat{p}_*(x) = \frac{f_*(x)}{n_*} \tag{1}$$

と定義される．x は離散要素である．$f_*(x)$, $* \in \{\mathrm{de}, \mathrm{nu}\}$ は密度 $p_*(x)$ に従う確率分布から得た要素 x の頻度であり，$n_* = \sum_x f_*(x)$ である．素朴な推定法は上式のように，確率分布を相対頻度 $\widehat{p}_*(x)$ で求めて比を取る方法である．そして，保守的な推定法 $\widehat{r}(x)$ は

$$\widehat{r}(x) = \left(\frac{f_{\mathrm{de}}(x)}{n_{\mathrm{de}}} + \lambda \right)^{-1} \frac{f_{\mathrm{nu}}(x)}{n_{\mathrm{nu}}}$$

と定義される．ここで λ (> 0) は正則化パラメータであり，これが頻度に応じて尤度比を保守的に見積もる効果を生む．具体的な推定例を用いて，先述した第一の問題を示し，$\widehat{r}(x)$ の作用も確認する．離散要素 x_1 から x_3 について，表 1 に示す頻度が得られたとする．まず x_1 と x_2 に着目すると，これらは頻度が大きく異なるが，$r_{\mathrm{MLE}}(x_1)$ と $r_{\mathrm{MLE}}(x_2)$ はともに 50 と等しく大きな値になった．ここで $f_{\mathrm{nu}}(x_1) = 1$ は偶然に出現した可能性がある．したがって，$r(a_k)$ の推定に $r_{\mathrm{MLE}}(x)$ を利用すると，偶然に出現した誤った属性値の尤度比に過大な推定値を与えることがある．一方で，$\widehat{r}(x_1)$ と $\widehat{r}(x_2)$ を見ると，頻度に応じて推定値を 50 よりも保守的に見積もることが分かる（それぞれ 47.6，8.3）．次に x_2 と x_3 に着目すると，$f_{\mathrm{nu}}(x_2)$ と $f_{\mathrm{nu}}(x_3)$ は 1 しか差がないが，$r_{\mathrm{MLE}}(x_2)$ と $r_{\mathrm{MLE}}(x_3)$ は大きく変動しており（それぞれ 50，100），推定値が安定していない．対して，$\widehat{r}(x_2)$ と $\widehat{r}(x_3)$ は変化が少なく（それぞれ 8.3，16.7），$\widehat{r}(x)$ は頻度の変動に対してロバストな推定法であることが分かる．分類タスクに尤度比を利用するとき，$r(a_k)$ の推定に保守的な推定法を用いると有益なケースがある．

第二の問題は，インスタンスを確率的に扱う上でよく知られた問題である．本稿では，この問題解決のために開発された特徴重み付け法を次のように組み合わせる．

$$r(y) \approx \prod_{k=1}^{n} r(a_k)^{w_k}$$

w_k は属性 A_k に対する重みである．A_k が有用であれば w_k は大きくなり，$\widehat{r}(a_k)$ が $r(y)$ の推定に与える影響は大きくなる．逆のケースでは w_k は小さくなり，$\widehat{r}(a_k)$ が $r(y)$ の推定に与える影響は小さくなる．本稿では重みの計算に，Correlation-based Feature Weighting (CFW) filter [3] という手法を用いる．これまでに我々は，自然言語処理における N-gram に対して，同様の尤度比推定法を提案した [4]．本稿では，インスタンスの尤度比推定に向けて，その推定法を再定義する．

人工的に生成した実験データを用い，尤度比による二値分類の実験を行う．表 1 の推定例による知見から，a_k の頻度および訓練データにおける y のクラス比（n_{de} と n_{nu} の比）が保守的な推定の有効性に影響を及ぼすと判断し，これらの条件を変化させて実験する．また，属性の有用性を変化させる実験も行い，尤度比推定における CFW の有効性を検証する．実験の結果，分類を誤らせる低頻度の属性値がデータに混入し，クラス比が大きく偏る状況（$n_{\mathrm{de}} >> n_{\mathrm{nu}}$）で，保守的な推定法が特に効果的であった．また，CFW の有効性も確認できた．

2 関連研究

確率推定を行い，その比を取って尤度比を推定するアプローチは，大きな推定誤差を生むことが知られている [5]．そこで，確率推定せずに尤度比を直に推定する “直接推定法” が提案されてきた [6, 7, 8, 9]．これらの直接推定法は，実数などの連続値をもとに，連続空間上で定義される尤度比を推定する．対して本稿では，離散値の観測頻度を用いて，離散空間上で定義される尤度比を推定する．そこで我々は，菊地らによる推定法 [2] を利用する．この推定法では，直接推定法の一つである拘束無し最小二乗重要度適合法（uLSIF） [9] に変更を加え，離散空間上の尤度比を推定できるようにした．菊地らの推定法では，最適化の過程で導入する正則化により，頻度に応じて尤度比を保守的に推定できる．

必要な属性をその度合いに応じて強調し，不要な属性の悪影響を低減する手段として，特徴重み付け法が提案されている．重み付けのタイプは重みの計算方法により，フィルタ法 [3] とラッパー法 [10] の二つに大きく分けられる．フィルタ法は訓練を行う前に，訓練データの情報を手がかりに各属性の重みを決定する．ラッパー法は，分類精度といった結果のフィードバックを手がかりに，最適化な重みを逐次的に決定する．本稿では，重み付けする尤度比推定量がハイパーパラメータを持つため，計算量が軽いフィルタ方式の重み付け法 CFW [3] を採用する．特徴重み付け法と類似したアプローチとして特徴選択法が提案されている [11]．特徴選択法は属性に 0 または 1 の重みを与え，有用な属性のみを選ぶ手法である．特徴選択法では用いる属性を削減するため，分類，回帰，推薦といった実用上のタスクを効率化できる．一方で，特徴重み付け法は属性の重みを柔軟に調節し，実用タスクを精度よく行える．

人工的に生成したデータはよく利用される．例えば，個人情報を含む実データを扱うにはプライバシーの懸念があるため，実データを模倣した人工データを用いる場合がある [12]．また人工データを用いた実験では，データ量やデータの性質といった実験条件を自由に制御でき，手法の長所や短所の検証が容易になる．実データから人工データを生成するソフトウェアもあり，代表的なものとして Synthetic Data Vault [13] と DataSynthesizer [14] がある．両方とも，インスタンス集合を入力として，同様の傾向を持つ人工的なインスタンス集合を出力する．本稿では，元のインスタンスの傾向をよりよく再現できる DataSynthesizer を用いる [15]．あるアルゴリズムに基づいて人工データを生成することもある [16]．しかし我々の調査した限り，属性値の種類数や出現頻度を制御できるデータ生成法は見つからなかった．ゆえに本稿では，Iris データセット[1)]をもとに人工データを生成し，連続的な属性値を異なるビン数で離散化することで，属性値の種類数や頻度を変化させる．

3 前提知識

提案法の説明に必要となる，尤度比の保守的な推定法 [2] および特徴重み付け法 [3] について述べる．

3.1 尤度比の保守的な推定法

尤度比推定の問題設定を説明する．あるデータが含む離散要素 x の集合を $D \subset \mathbb{U}^v$ とする．$\mathbb{U}^v$ は存在しうる v 種類の要素からなる集合で，情報理論では有限アルファベットと呼ばれる．確率密度関数 $p_{\mathrm{de}}(x)$，$p_{\mathrm{nu}}(x)$ を持つ確率分布に従う二つの i.i.d. 標本

$$\{x_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{de}}(x), \quad \{x_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{nu}}(x)$$

を得たとする．“de” と “nu” はそれぞれ，尤度比の分母と分子を表す添え字である．$p_{\mathrm{de}}(x)$ が次の条件

$$p_{\mathrm{de}}(x) > 0 \quad \text{for all } x \in D$$

を満たすと仮定する．この仮定により，全ての x に対して尤度比の定義が可能になる．本節では，二つの標本 $\{x_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$，$\{x_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ から次の尤度比を推定する．

$$r(x) = \frac{p_{\mathrm{nu}}(x)}{p_{\mathrm{de}}(x)}$$

保守的な推定法 [2] は，尤度比の直接推定法である拘束無し最小二乗重要度適合法 （uLSIF） [9] をもとに導出された．本稿では，評価実験に焦点を当てるため，導出過程を省略する．最終的な推定量は

$$\widehat{r}(x) = \left(\frac{f_{\mathrm{de}}(x)}{n_{\mathrm{de}}} + \lambda\right)^{-1} \times \frac{f_{\mathrm{nu}}(x)}{n_{\mathrm{nu}}} \tag{2}$$

と定義される．この推定量は，推定モデルと真の尤度比との二乗誤差を最小化する問題の解として導出される．この問題で導入される正則化パラメータ λ (> 0) が，頻度に応じて $r(x)$ を保守的に見積もる．$\lambda = 0$ のとき，この式は尤度比の素朴な推定法（すなわち，$r(x)$ の分母・分子にあたる確率分布をそれぞれ相対頻度で推定し，その比を取る方法）と等しくなる．

3.2 特徴重み付け法 CFW

特徴重み付け法はナイーブベイズ分類器に対してよく用いられる．ナイーブベイズ分類器では，ある属性 A_k がクラス $c \in C$ に属する条件下で，他の属性と統計的

1） https://archive.ics.uci.edu/ml/datasets/iris

独立という条件付き独立性を仮定する．この仮定により，$p(y \mid c)$ を $p(a_k \mid c)$ の積として表現でき，y が訓練データに出現しなくても，それぞれの属性値 a_k が出現していれば $p(y \mid c)$ を推定できる．しかし，先述の条件付き独立性は実際には成立しないことが多く，属性間には何らかの依存関係がある場合がほとんどである．また，分類に無関係な属性や分類を誤らせる属性があると，$p(a_k \mid c)$ の積を取るアプローチでは分類精度が低下してしまう．そこで，分類における A_k の重要度に応じて $p(a_k \mid c)$ を重み付けする．重みを導入した Feature Weighted Naive Bayes は

$$\widehat{c}(y) = \arg\max_{c \in C} p(c) \prod_{k=1}^{n} p(a_k \mid c)^{w_k}$$

と定義される．ここで，$\widehat{c}(y)$ は y が属すると分類器が判定したクラスを表す．w_k は属性 A_k に対する重みである．A_k が y の分類において重要であれば w_k は大きくなり，そうでなければ w_k は小さくなる．w_k を計算する手法を総称して特徴重み付け法と呼ぶ．

本稿では，特徴重み付けの一つである Correlation-based Feature Weighting (CFW) filter [3] を用いる．CFW は，分類に寄与する重要な属性 A_k はクラスと強く相関し，他の属性 $A_{k'}$ $(k \neq k')$ と相関しないという考え方に基づく．この考え方のもとで，属性—クラス間の相関，属性間の相互相関を求め，その差をシグモイド変換したものを重みとして算出する．

CFW は相関を測る尺度として相互情報量を利用する．A_k に対する属性—クラス間の相互情報量を

$$I(A_k; C) = \sum_{a_k} \sum_{c} p(a_k, c) \log \frac{p(a_k, c)}{p(a_k)p(c)}$$

と定義する．同様に，A_k と $A_{k'}$ $(k \neq k')$ に対する属性間の相互情報量を

$$I(A_k; A_{k'}) = \sum_{a_k} \sum_{a_{k'}} p(a_k, a_{k'}) \log \frac{p(a_k, a_{k'})}{p(a_k)p(a_{k'})}$$

と定義する．なお，$I(A_k; C)$ と $I(A_k; A_{k'})$ の定義中にある各確率は，相対頻度を用いて推定する．そして，$I(A_k; C)$ と $I(A_k; A_{k'})$ をぞれぞれ

$$NI(A_k; C) = \frac{I(A_k; C)}{\frac{1}{n}\sum_{m=1}^{n} I(A_m; C)}$$

$$NI(A_k; A_{k'}) = \frac{I(A_k; A_{k'})}{\frac{1}{n(n-1)}\sum_{m=1}^{n}\sum_{m'=1 \wedge m' \neq m}^{n} I(A_m; A_{m'})}$$

として正規化する．分類に寄与する属性はクラスと強く相関し，他の属性とは相関しないという考え方に基づき，A_k に対する重み d_k を

$$d_k = NI(A_k; C) - \frac{1}{n-1} \sum_{k'=1}^{n} NI(A_k; A_{k'})$$

として計算する．その後，開区間 $(0, 1)$ で重みが値を取るように d_k をシグモイド変換する．以上より，A_k に対する最終的な重み w_k は

$$w_k = \frac{1}{1 + \mathrm{e}^{-d_k}}$$

となる．w_k が 1 に近づくと $p(a_k \mid c)^{w_k}$ は $p(a_k \mid c)$ に近づき，$p(a_k \mid c)$ が分類に及ぼす影響は大きくなる．それに対して，w_k が 0 に近づくと $p(a_k \mid c)^{w_k}$ は 1 に近づき，$p(a_k \mid c)$ が分類に及ぼす影響は小さくなる．

CFW は各属性とクラス変数 C が与えられれば利用でき，計算される重み w_k は 0 から 1 の範囲に収まる．このことから，重み付けする対象が条件付き確率ではなく尤度比であっても効果的であると着想した．

4 提案法

インスタンス $y = \langle a_1, a_2, \ldots, a_n \rangle$ に対する次の尤度比 $r(y)$ を推定する．A_k $(k = 1, 2, \ldots, n)$ は離散値属性とし，$a_k \in A_k$ は出現頻度が数え上げ可能な属性値とする．

$$r(y) = \frac{p_{\mathrm{nu}}(y)}{p_{\mathrm{de}}(y)}$$

a_k が他の属性値 $a_{k'}$ $(k \neq k')$ と統計的独立と仮定し，$r(y)$ を a_k に対する尤度比 $r(a_k)$ の積

$$r(y) = \prod_{k=1}^{n} r(a_k)$$

で表現する．これにより，a_k の頻度で $r(y)$ を推定できる．しかし，この方法には二つの問題がある．第一に，式 (1) のように $r(a_k)$ を相対頻度で単純に推定すると，低頻度の a_k に対する尤度比を過大推定することがある．$r(y)$ は $r(a_k)$ の積であるから，$r(a_k)$ の過大推定は $r(y)$ の推定値に悪影響を与えてしまう．第二に，a_k 間の統計的独立は実際には成立しないことが多い．加えて，分類に無関係な属性や分類を誤らせる属性があると，$r(y)$ の推定値に悪影響を及ぼす．

第一の問題には，3.1 節の式 (2) で示した保守的な推定法を，$r(a_k)$ の推定に用いて対処する．第二の問題には，3.2 節の CFW によって計算した重み w_k を $r(a_k)$ へと導入して対処する．CFW を用いる際は，密度 $p_{\mathrm{de}}(a_k)$ を持つ確率分布から得た a_k にクラス c_{de} を，密度 $p_{\mathrm{nu}}(a_k)$ を持つ確率分布から得た a_k にクラス c_{nu} をそれぞれ割り当てる．また a_k の頻度をそのまま用いると，$f_{\mathrm{nu}}(a_k) = 0$ となる a_k が一つでもあった場合に，$r(y)$ の推定値がゼロになってしまう．そこで頻度に微小な補正を加える．以上より，本稿で提案する推定法は

$$r_{\mathrm{ours}}(y) = \prod_{k=1}^{n} \widetilde{r}(a_k)^{w_k} \tag{3}$$

$$\widetilde{r}(a_k) = \left(\frac{f_{\mathrm{de}}(a_k) + 1}{n_{\mathrm{de}} + 2} + \lambda \right)^{-1} \times \frac{f_{\mathrm{nu}}(a_k) + 1}{n_{\mathrm{nu}} + 2}$$

と定義される．補正頻度による確率推定量 $\frac{f_*(a_k)+1}{n_*+2}$, $* \in \{\mathrm{de}, \mathrm{nu}\}$ は，密度 $p_*(a_k)$ を持つ確率分布を a_k が出現するか否かのベルヌーイ試行による確率分布と仮定し，事前分布を一様分布としたときの事後期待値と一致する．確率推定に自然言語処理におけるスムージング法 [17] を用いる余地もあるが，一般にスムージング法による確率推定値は相対頻度よりもはるかに低く見積もられる．この場合，スムージング法による確率推定値を尤度比推定に利用すると，過大推定につながることが明らかになっている [2]．ゆえに本稿では，頻度に最低限の微小

な補正を加える．なお，頻度の補正はゼロ頻度を低頻度へと変えるため，過大推定の原因となる可能性がある．しかし上式では，正則化パラメータ λ (> 0) の作用で過大推定を抑制できる．なお本手法は，我々の先行研究による推定法 [4] をインスタンスの尤度比推定に向けて再定義したものである．

5 評価実験

離散値属性を持つインスタンスの分類問題を解き，提案法が有効な状況を明らかにする．そのために，人工データに基づく三つの実験を行う．人工データを用いることで，実験条件を系統的に変化させ，提案法の長所と短所を探る．第一の実験では，訓練データにおける属性値の出現頻度，インスタンスのクラス比を変化させる．第二の実験では，訓練データで未観測の属性値を持つインスタンスを評価データに混入させる．第三の実験では，属性値の出現傾向が正規分布に従う人工的な属性 A_k^+ をインスタンスに加える．そして，正規分布の標準偏差を変えることで，分類に対する “属性値の有用性” を調節し，追加属性が尤度比ベースの分類精度に与える影響を調べる．第一と第二の実験は，尤度比の保守的な推定法が有効な状況を調査するために行う．第三の実験は特徴重み付け法の有効性検証のために行う．

5.1 Iris データセット

実験で用いる人工データは，分類問題のベンチマークデータセットとして有名な Iris データセット[1)]から作成する．本節では，Iris データセットについて述べ，実験データを得るためのデータ源 S の作成方法を説明する．Iris データセットは 3 クラス 50 インスタンスずつの合計 150 インスタンスを含む．インスタンスは 4 つの測定値（属性値）a_1〜a_4 を持つ．

- a_1：sepal.length（がく片の長さ）
- a_2：sepal.width（がく片の幅）
- a_3：petal.length（花びらの長さ）
- a_4：petal.width（花びらの幅）

そして，3 種類の花の名前 Setosa，Versicolor，Virginica が正解（クラスラベル）として付与されている．以降，3 つのクラスをそれぞれ c_{se}，c_{ve}，c_{vi} と表記する．属性値は実数（連続値）であり，実験では属性値にビニング処理を適用して連続値を離散化する．元々が離散的な属性値の頻度を制御することは難しいが，ビニング処理では区切るビンの数を変えることで，属性値の種類数と出現頻度を変更できる．具体的なビン数は実験の説明で述べる．また，Iris データセットは分類が容易で，ナイーブベイズ分類器を用いても分類精度は 90%を超える．この理由として，属性値の取りうる値の範囲がクラスごとによく分かれていることが挙げられる．そこで，訓練データを 2 クラスのみのインスタンスで構成し，評価データには訓練データに存在しない残り 1 クラスのインスタンスも加えることで，未知の属性値を持つデータが混入した状況を容易に再現できる．

ただし，Iris データセットは実験にそのまま用いるには小規模である．そこで，Iris データセットの傾向を模倣したデータセットを生成し，実験データを

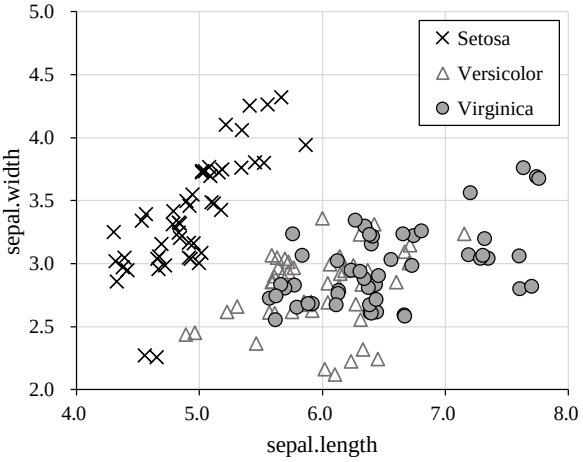


図 1　がく片の長さ a_1 と幅 a_2 の分布

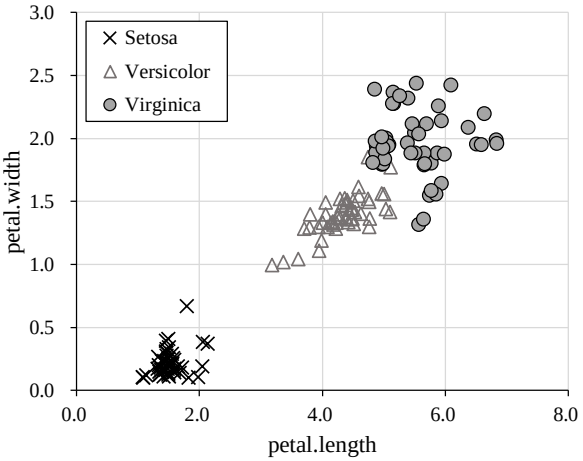


図 2　花びらの長さ a_3 と幅 a_4 の分布

得るためのデータ源 S とした．S を生成するために，DataSynthesizer [14] を correlated attribute mode で用いた．なお他のデータ生成ツールと比較して，DataSynthesizer は元のデータセットの傾向をよりよく再現できる [15]．このモードではベイジアンネットワークを構築し，相関する属性をモデル化する．そして，Iris データセットと類似した傾向を持つインスタンス集合を出力する．生成するインスタンス数を 50,000 件とし，元のデータセットの傾向を忠実に再現するため，属性値には差分プライバシーのためのノイズを含めないようにした．他のパラメータはデフォルトのまま用いた．S から各クラスのインスタンスを 50 件ずつランダムに抽出し，属性値間の分布をプロットしたものを図 1 と図 2 に示す．これらの図から，Setosa のインスタンスは，他 2 種類のインスタンスに比べて分類しやすいことが分かる．また，Versicolor と Virginica はがく片の測定値では分類しにくいが，花びらの測定値では分類しやすいことも分かる．

5.2 実験 1：属性値の頻度などに対する頑強性

この実験では，訓練データ中の属性値 a_k の出現頻度，インスタンスのクラス比を変化させて，提案法のふるまいを観察する．尤度比は二値分類のための統計量である

から，実験では c_{ve} と c_{vi} のインスタンスのみを用い，それらを正しく二値分類する問題を解く．まずは次の実験条件から一つずつを選択する．

ビン数　100，1,000
訓練データ中のクラス比　$c_{vi} : c_{ve} = 1 : 1$，$1 : 9$

次に，5.1 節で述べたデータ源 S のインスタンスが含む各属性値を，指定したビン数でビニング処理して離散化する．ビニング処理では，各属性値の最大値と最小値の間を等間隔で分割する．S から実験データをサンプリングする．訓練データが 5,000 件，評価データが 1,000 件となるように，S にある 50,000 件からインスタンスをランダムに非復元抽出する．評価データのクラス比は常に $c_{vi} : c_{ve} = 1 : 1$ とする．すなわち，評価データは c_{vi} と c_{ve} のインスタンス 500 件ずつを含む．訓練データのクラス比は上記の条件で変化させる．$c_{vi} : c_{ve} = 1 : 1$ の場合は，c_{vi}，c_{ve} のインスタンス 2,500 件ずつが訓練データに含まれる．$c_{vi} : c_{ve} = 1 : 9$ の場合は，c_{vi} のインスタンス 500 件，c_{ve} のインスタンス 4,500 件が訓練データに含まれる．なお，条件を変えるごとに，上記の手順で実験データの作成をやり直す．

実験手順を述べる．a_k の出現頻度を訓練データから計数する．重み w_k もここで計算する．評価データの全インスタンスに対し，次の尤度比を推定する．

$$r(y) = \frac{p(y \mid c_{vi})}{p(y \mid \bar{c}_{vi})} \tag{4}$$

$\bar{c}_{vi}$ は c_{vi} 以外のクラスであり，本実験では c_{ve} と等しい．$r(y)$ が高いほど y は c_{vi} に属しやすく，$r(y)$ が低いほど y は c_{ve} に属しやすい．我々は尤度比を二値分類の順位付け関数として用い，順位付けの良し悪しから推定法の有効性を判断する．尤度比の推定法に対する評価手順は次の通りである．まず，推定法ごとに評価データの全インスタンスを推定値の降順に並び替える．そして二値分類の正誤判定を行う．y が c_{vi} に属していれば正解，$\bar{c}_{vi}$（$= c_{ve}$）に属していれば不正解とする．最後に，正誤判定の結果を用いて ROC 曲線を描き，曲線下面積（ROC-AUC）を算出する．AUC が大きい推定法ほど，二値分類の観点から優れていると判断する．

保守的な推定法，特徴重み付け法 CFW，およびそれらの組み合わせの有効性を検証するため，以下の推定法を比較手法として用いる．

手法 1：ベースライン　保守的な推定法と CFW の両方を用いない．この推定法は式 (3) の正則化パラメータ λ を 0，重み w_k を 1 とした場合と等しい．

手法 2：重み付け　CFW のみ用いる．この推定法は式 (3) の λ を 0 とした場合と等しい．

手法 3：保守的　保守的な推定法のみ用いる．この推定法は式 (3) の w_k を 1 とした場合と等しい．本実験では，λ を 10^{-4}，10^{-3}，10^{-2} と設定した際のふるまいを比較する．

手法 4：保守的 ＋ 重み付け（提案法）　保守的な推定法と CFW の両方を用いる．この推定法は式 (3) で定義される．λ を 10^{-4}，10^{-3}，10^{-2} と設定した際のふるまいを比較する．

表 2　ビン数を 100 としたときの ROC-AUC

尤度比の推定法	$c_{vi} : c_{ve}$ 1 : 1	1 : 9
ベースライン	0.991	0.986
重み付け	0.992	0.989
保守的（$\lambda = 10^{-4}$）	0.990	0.986
保守的（$\lambda = 10^{-3}$）	0.990	0.985
保守的（$\lambda = 10^{-2}$）	0.987	0.979
保守的（$\lambda = 10^{-4}$）＋ 重み付け	0.992	0.989
保守的（$\lambda = 10^{-3}$）＋ 重み付け	0.992	0.988
保守的（$\lambda = 10^{-2}$）＋ 重み付け	0.990	0.984

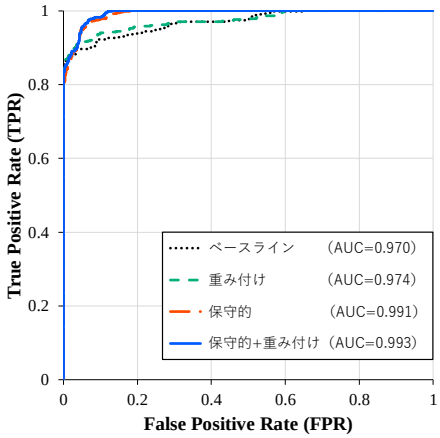
表 3　ビン数を 1,000 としたときの ROC-AUC

尤度比の推定法	$c_{vi} : c_{ve}$ 1 : 1	1 : 9
ベースライン	0.984	0.981
重み付け	0.988	0.986
保守的（$\lambda = 10^{-4}$）	0.983	0.981
保守的（$\lambda = 10^{-3}$）	0.979	0.975
保守的（$\lambda = 10^{-2}$）	0.959	0.926
保守的（$\lambda = 10^{-4}$）＋ 重み付け	0.987	0.986
保守的（$\lambda = 10^{-3}$）＋ 重み付け	0.983	0.982
保守的（$\lambda = 10^{-2}$）＋ 重み付け	0.968	0.943

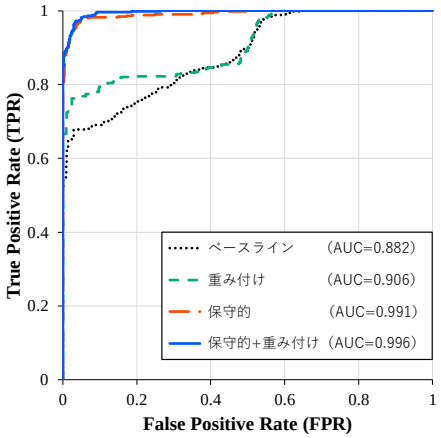
実験 1 の結果について述べる．ビン数 100，1,000 のときの ROC-AUC を表 2 と表 3 に示す．どの実験条件でも全手法の AUC が 0.9 を上回る．しかし，保守的な推定を行う推定法は，“ベースライン” よりも AUC が低いケースが多い．全手法で分類が容易な理由として，属性値が特定の範囲に収まりやすく，クラスごとに値の範囲が区別できる傾向にあることが考えられる．そのような分類に適したデータでは，ビン数を増やし訓練データのクラス比を極端にしても，評価データにある属性値は訓練データにも含まれやすい．さらに外れ値のような属性値も少なく，属性値が低頻度でも分類に悪影響があまりないと考えらえる．この状況下で保守的な推定は必要はなく，保守的に推定する度合い（すなわち λ の値）を大きくすると AUC が下がると考える．一方で CFW を利用すると AUC はわずかに向上する．CFW で計算した重みを表 4 に示す．この表と 5.1 節の図 1，図 2 で示した属性値の分布を比較すると，分類の有用性が低いがく片の属性には小さい重みが，有用性が高い花びらの属性には大きい重みが付与された．よってわずかだが，尤度比推定における CFW の有効性が示唆された．

表 4　各属性に対する重み（実験 1）

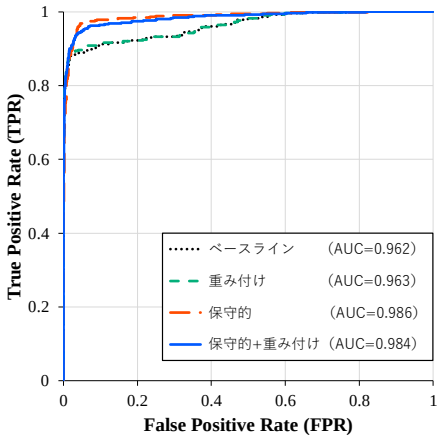
ビン数	$c_{vi} : c_{ve}$	がく片 w_1	w_2	花びら w_3	w_4
100	1 : 1	0.387	0.368	0.608	0.637
	1 : 9	0.386	0.342	0.631	0.641
1,000	1 : 1	0.420	0.365	0.600	0.616
	1 : 9	0.429	0.370	0.580	0.621



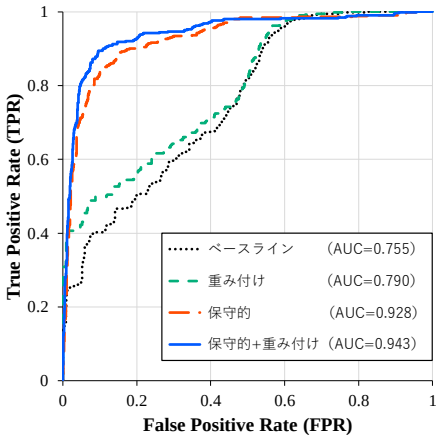
(a) ビン数 100，$c_{\mathrm{vi}} : c_{\mathrm{ve}} = 1 : 1$



(b) ビン数 100，$c_{\mathrm{vi}} : c_{\mathrm{ve}} = 1 : 9$



(c) ビン数 1,000，$c_{\mathrm{vi}} : c_{\mathrm{ve}} = 1 : 1$



(d) ビン数 1,000，$c_{\mathrm{vi}} : c_{\mathrm{ve}} = 1 : 9$

図 3　評価データに c_{se} のインスタンスを混入させたときの ROC 曲線と ROC-AUC

5.3　実験 2：未知インスタンスに対する頑強性

実験 1 の結果から，未知の属性値や外れ値が少ない分類に適したデータでは，保守的な推定が不要なことが示唆された．そこで本実験では，訓練データで観測されない属性値を持ち，分類に無関係なインスタンスを評価データへ混入させる．そして，未知のインスタンスに対する各推定法のふるまいを明らかにする．

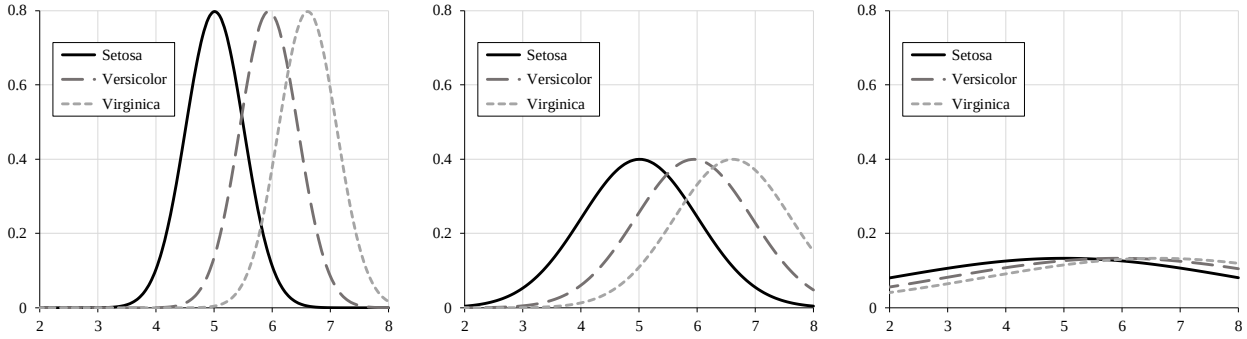
実験 1 と同様に，離散化のビン数，訓練データ中のクラス比 $c_{\mathrm{vi}} : c_{\mathrm{ve}}$ を変化させる．実験 1 と異なるのは実験データである．訓練データには c_{vi} と c_{ve} のインスタンスのみを含むのに対し，評価データには c_{vi} と c_{ve} のインスタンス 500 件ずつに加え，Setosa のクラス c_{se} に属するインスタンス 500 件も混入させる．図 1 と図 2 で示したように，c_{se} の属性値は他クラスの属性値と存在範囲が重複しない傾向にある．このことから，c_{se} の情報が訓練データにないとき，c_{se} のインスタンスは訓練データにとって未知の属性値を含みやすい．推定する尤度比は式 (4) で定義されるが，本実験での $\bar{c}_{\mathrm{vi}}$ は $c_{\mathrm{se}} \vee c_{\mathrm{ve}}$ であることに注意する．実験手順および比較手法も実験 1 と同様である．なお，正則化パラメータを持つ手法 3 と手法 4 については，ROC-AUC が最大となった $\lambda = 10^{-2}$ の結果のみを掲載して議論する．

表 5　訓練データにない属性値の数（ビン数 1,000）

$c_{\mathrm{vi}} : c_{\mathrm{ve}}$	クラス	属性値			
		a_1	a_2	a_3	a_4
1 : 1	c_{se}	207	140	500	500
	c_{ve}	12	9	12	5
	c_{vi}	10	9	7	7
1 : 9	c_{se}	170	224	500	500
	c_{ve}	10	2	2	2
	c_{vi}	68	29	99	79

実験 2 の結果について述べる．評価データに c_{se} のインスタンスを混入させたときの ROC 曲線と ROC-AUC を図 3 に示す．実験 1 と異なり，保守的な推定の有無で手法のふるまいに大きな違いが観測された．表 2，表 3 と AUC の値を見比べると，保守的な推定をしない “ベースライン”，“重み付け” では AUC の値が大きく低下したことが分かる．特に，ビン数が多く訓練データのクラス比 $c_{\mathrm{vi}} : c_{\mathrm{ve}}$ が 1 : 9 と偏る図 3(d) で，AUC の低下が著しい．それに対して保守的な推定を用いた手法は，そのような場合でも高い AUC を維持している．

尤度比の推定値と分類精度の関係を考察する．ビン数



(a) $\sigma = 0.5$　(b) $\sigma = 1.0$　(c) $\sigma = 3.0$

図 4　A_1^+ に対する正規分布．σ が小さいほど，分類に有用な属性値が生成されやすい．

表 6　$r(y)$ に対する推定値の平均（ビン数 1,000）

$c_{\mathrm{vi}} : c_{\mathrm{ve}}$	クラス	尤度比の推定法	
		ベースライン	保守的＋重み付け
1 : 1	c_{se}	1.412	4.113×10^{-3}
	c_{ve}	0.533	2.974×10^{-2}
	c_{vi}	228.356	22.405
1 : 9	c_{se}	2,069.414	4.247×10^{-2}
	c_{ve}	13.51	4.771×10^{-2}
	c_{vi}	9,203.913	0.170

1,000 の条件下で訓練データにはない属性値の個数を表 5 にまとめた．訓練データは c_{se} のインスタンスを含まないため，c_{se} の属性値はゼロ頻度のものが多い．c_{se} の花びらに関する属性値は訓練データに一つもないことが分かる．“ベースライン” と “保守的＋重み付け”（提案法）に絞り，各クラスのインスタンスに対する尤度比推定値の平均を表 6 にまとめた．$c_{\mathrm{vi}} : c_{\mathrm{ve}} = 1 : 1$ のときは，“ベースライン” でも c_{vi} に対する尤度比のみを，（他クラスと相対比較して）高めに推定できた．しかし，$c_{\mathrm{vi}} : c_{\mathrm{ve}} = 1 : 9$ では保守的な推定をしないと，訓練データにない属性値の尤度比を高めに見積もる．結果として “ベースライン” では，c_{se} に対する推定値が高くなり，AUC の低下を招いたと考える．一方で保守的な推定を用いた手法は，c_{vi} のインスタンスのみを高めに推定できる．以上のことから，訓練データ中のクラス比が極端かつゼロ頻度や低頻度の属性値を多く扱う状況で，保守的な推定法が特に有効であると考える．

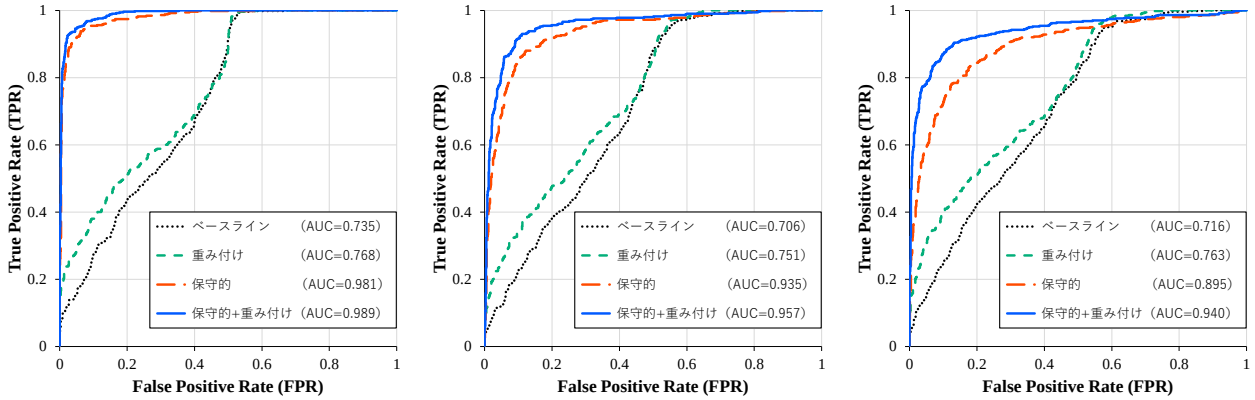
5.4　実験 3：追加属性による CFW の有効性検証

実験 2 では，保守的な推定法が有効な状況を明らかにしたが，特徴重み付け法 CFW の有無に対する優劣の差が小さい．そこで本実験では，尤度比推定における CFW の有効性を検証する．そのために，分類に対する “有用性” を調節できる属性値を人工的に生成し，インスタンスに追加する．悪影響のある（つまり分類を誤らせるような）属性値が追加された際は，重み付けの有無に対する性能差が明瞭になると予想される．

本実験では，属性 A_1〜A_4 をもとに作成した属性値 a_1^+〜a_4^+ を，S の各インスタンス y へ追加し，$y^+ = \langle a_1, a_2, a_3, a_4, a_1^+, a_2^+, a_3^+, a_4^+ \rangle$ とする．$a_k^+ \in A_k^+$ は正規分布 $\mathcal{N}(\mu_{k,c(y)}, \sigma)$ からサンプリングされる．ここで $\mu_{k,c(y)}$ はインスタンス y が属するクラス $c(y)$ における属性値 $a_k \in A_k$ の平均を表す．σ は正規分布の標準偏差であり，実験条件の一つとする．本実験では $\sigma = 0.5, 1.0, 3.0$ から一つを選択する．なお σ の値は $c(y)$ と k によらず一定である．A_1（sepal.length）をもとに描いた正規分布を図 4 に示す．例えば a_1^+ を付与するインスタンスがクラス c_{se} に属していれば，Setosa の正規分布（図中の黒色実線）から a_1^+ がサンプリングされる．σ が小さい場合は，正規分布の間の重複範囲が狭く，識別力のある有用な属性値が得られる．一方で σ が大きい場合は，それぞれの正規分布が広く重複しており，分類を誤らせる属性値が得られる．

実験条件として $\sigma = 0.5, 1.0, 3.0$ から一つを選択する．その他の条件は，実験 2 で保守的な推定が有効であったビン数 1,000，クラス比 $c_{\mathrm{vi}} : c_{\mathrm{ve}} = 1 : 9$ に固定する．y^+ からなるデータ源 S^+ からインスタンスを抽出し，実験データを作成する．評価データには実験 2 と同様，3 クラスのインスタンスを含む．なお，σ を変えるごとに，実験データの作成をやり直す．手法 3 と手法 4 については，ROC-AUC が最大となった $\lambda = 10^{-2}$ の実験結果のみを掲載して議論する．

実験 3 の結果について述べる．追加属性を導入したときの ROC 曲線と ROC-AUC を図 5 に示す．まず図 5(a) に示した $\sigma = 0.5$ の ROC 曲線に着目する．保守的な推定の有無によって，AUC の値に大きな差がある．追加属性のない結果である図 3(d) とも比較すると，σ が 0.5 と小さい場合でも保守的な推定を使用しない 2 手法（“ベースライン”，“重み付け”）は，AUC が低下した．$\sigma = 0.5$ でも正規分布の間に重複部分があり，分類を誤らせる属性値が生成されることがある．これらの 2 手法は，そのような少量の属性値による悪影響を強く受けると考えられる．一方で保守的な推定を使用する 2 手法は，AUC が向上しており，属性値を有効に扱えることが示唆された．次に図 5(b) と図 5(c) に示す $\sigma = 1.0$，3.0 の ROC 曲線に着目する．ここでも保守的な推定の有無で AUC の値に大きな差がある．“保守的” と “保守的＋重み付け” の ROC 曲線を見ると，σ が大きくなるほど 2 手法間の差が開く．表 7 に示す属性の重みにも着目すると，σ が 0.5 の場合は追加属性 A_k^+ の重みが本来の属性 A_k の重みと類似するのに対して，σ が 3.0 の場合は追加属性 A_k の重みが一律に小さい．したがって，尤度比推定でも特徴重み付けは有効なことが示唆された．



(a) $\sigma = 0.5$　(b) $\sigma = 1.0$　(c) $\sigma = 3.0$

図 5　追加属性 A_k^+ を導入したときの ROC 曲線と ROC-AUC．ビン数 1,000，$c_{\mathrm{vi}} : c_{\mathrm{ve}} = 1 : 9$ である．

表 7　各属性に対する重み（実験 3）

σ	A_1〜A_4 に対する重み				A_1^+〜A_4^+ に対する重み			
	w_1	w_2	w_3	w_4	w_1^+	w_2^+	w_3^+	w_4^+
0.5	0.454	0.384	0.642	0.672	0.441	0.353	0.610	0.440
1.0	0.509	0.403	0.714	0.743	0.384	0.368	0.457	0.394
3.0	0.515	0.416	0.739	0.777	0.379	0.378	0.377	0.374

おわりに

離散値属性を持つインスタンスの尤度比推定法を提案した．提案法は $r(y)$ を $r(a_k)$ の積で近似する際の問題点に着目し，二つの技術を取り入れた．第一に，低頻度から求まる $r(a_k)$ の推定値が過大になるのを防ぐため，尤度比の保守的な推定法を導入した．第二に，$r(y)$ の推定に不要な属性の悪影響を低減するため，特徴重み付け法 CFW を導入した．そして，Iris データをもとに作成した人工データを用いて二値分類を解き，提案法が有効な状況を調べた．結果として，分類を誤らせる属性値がデータに混入し，クラス比が大きく偏る状況で，保守的な推定法が特に効果的であり，CFW の有効性も確認した．図 5(a) より，“ベースライン” と比較して提案法を用いると ROC-AUC が最大 0.254 向上した．

謝辞

本研究の一部は JSPS 科研費 JP19K12266，JP22K18006 の助成を受けたものです．

参考文献

[1] K. Nakanishi, T. Tanaka, and N. Ueda. *Asymptotic properties of area under the ROC curve via likelihood ratio based ranking function*. IEICE Technical Report, 2015. IBISML2014-92.

[2] 菊地真人, 川上賢十, 吉田光男, 梅村恭司. 観測頻度に基づくゆう度比の保守的な直接推定. 電子情報通信学会論文誌 D, Vol. J102-D, No. 4, pp. 289–301, 2019.

[3] L. Jiang, L. Zhang, C. Li, and J. Wu. A correlation-based feature weighting filter for naive bayes. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 31, No. 2, pp. 201–213, 2018.

[4] 菊地真人, 吉田光男, 梅村恭司. 特徴重み付けを用いた低・ゼロ頻度 N-gram に対する尤度比の推定法. In *DEIM Forum 2020*, pp. 1–8, 2020.

[5] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and semiparametric models*. Springer Science & Business Media, 2012.

[6] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, pp. 601–608, 2007.

[7] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *ICML*, pp. 81–88, 2007.

[8] M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, pp. 1433–1440, 2008.

[9] T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, Vol. 10, pp. 1391–1445, July 2009.

[10] N. A. Zaidi, J. Cerquides, M. J. Carman, and G. I. Webb. Alleviating naive bayes attribute independence assumption by attribute weighting. *Journal of Machine Learning Research*, Vol. 14, No. 1, pp. 1947–1988, 2013.

[11] X. Deng, Y. Li, J. Weng, and J. Zhang. Feature selection for text classification: A review. *Multimedia Tools and Applications*, Vol. 78, No. 3, pp. 3797–3816, 2019.

[12] H. Surendra and H. S. Mohan. A review of synthetic data generation methods for privacy preserving data publishing. *International Journal of Scientific & Technology Research*, Vol. 6, No. 3, pp. 95–101, 2017.

[13] N. Patki, R. Wedge, and K. Veeramachaneni. The synthetic data vault. In *DSAA*, pp. 399–410, 2016.

[14] H. Ping, J. Stoyanovich, and B. Howe. DataSynthesizer: Privacy-preserving synthetic datasets. In *SSDBM*, pp. 1–5, 2017.

[15] M. Hittmeir, A. Ekelhart, and R. Mayer. On the utility of synthetic data: An empirical evaluation on machine learning tasks. In *ARES*, pp. 1–6, 2019.

[16] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos. A review of feature selection methods on synthetic data. *Knowledge and information systems*, Vol. 34, No. 3, pp. 483–519, 2013.

[17] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, Vol. 13, No. 4, pp. 359–394, 1999.