The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

二値意思決定のためのヒヤリハットを用いた不均衡判別学習

Imbalanced Classification with Near-misses for Binary Decision-making

谷本 啓 *1*2*3 Akira Tanimoto　山田 聡 *1 So Yamada　竹之内 高志 *2*4 Takashi Takenouchi　鹿島 久嗣 *2*3 Hisashi Kashima

*1NEC NEC　*2理化学研究所 Riken　*3京都大学 Kyoto University　*4公立はこだて未来大学 Future University Hakodate

We consider a prediction-based decision-making problem, in which a binary decision corresponds to whether or not a numerical variable is predicted to exceed a given threshold. The final goal is to predict a binary label, however, we can exploit the numerical variable in the training phase as side-information. In addition, we focus on class-imbalanced situation. We investigate on an idea of using near-miss samples, which is specified by the numerical variable, to deal with the class-imbalance. We present the benefit of exploiting the side-information theoretically as well as experimentally.

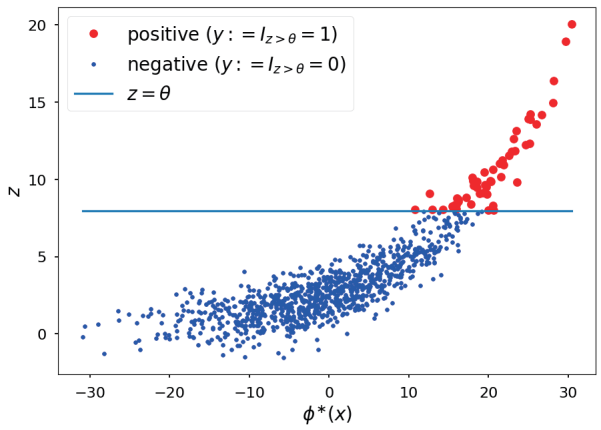

図 1: GLM モデルにより生成した人工データ．“正例度合い”z が訓練データ中に得られているが，最終的に得たい出力は y であるとき，z を不必要に予測することなく，しかし z の情報を利用して y の予測器を学習することを考える．

1 背景

我々は，不均衡データの判別問題であって，データに補助情報として “正例度合い” が含まれるようなものを考える．例えば水位情報に基づく洪水予測では，河川の水位が堤防の高さを超えた場合に洪水となり，その時点で避難を完了しているべき状況となる．よって，最終的に解きたい問題は河川水位が翌日や数時間後に堤防の高さを超えるかどうかという二値の予測である．しかし，これを通常の判別問題として解く場合，正例，すなわち実際に洪水が起こった事例数が相当程度必要となり，洪水予測のような安全にかかわる分野での適用は困難である．そこで我々は、河川水位が堤防の高さに近づいた事例を部分的に正例として活用することにより不均衡を改善することを考える．この場合，観測された河川水位を正例度合いと呼ぶ．

素朴なアプローチとしては，まず水位を予測し，その予測値に閾値を適用することによって避難すべきかどうかを判断するという方式が考えられる．しかし，あくまで最終的に行いたいのは避難すべきかどうかの判断だけであることを考えれば，中間問題である水位予測を解くことが悪影響する場合も考えられる．この方式については第 3.1節にて議論する．

一方で，正例数と負例数が不均衡なデータから判別器を学習することが難しいことはよく知られている．特に，正例が実際の事故や災害を意味するなど稀な現象と対応する場合，第 4章で解析するように，正例数がボトルネックとなる．

そこで我々はこのような，最終的に行いたい意思決定は二値であるが学習時には数値データが得られているような状況に対し，数値データを補助情報として扱うことによるサンプル効率的な判別学習を行う手法を提案する．理論解析及び実データでの実験によって既存の判別手法と提案手法を比較し，高度に不均衡な設定における提案手法の優位性を示す．

連絡先: 谷本 啓，ＮＥＣ，a-tanimoto@ay.jp.nec.com

2 問題設定

判別モデル $f : \mathcal{X} \to \mathcal{Y}$ の学習が我々の目的である．ここで，$\mathcal{X}$ は説明変数空間，$\mathcal{Y} = \{0, 1\}$ はラベル空間である．さらに，各ラベル $y_n \in \mathcal{Y}$ は次のように，正例度合い $z_n \in \mathbb{R}$ と与えられた閾値 θ から決定される．

$$y_n = \begin{cases} 1 & (z_n \geq \theta) \\ 0 & (z_n < \theta). \end{cases}$$

簡単のため，以後 $\theta = 0$ とする ($z_n - \theta$ をあらたな z_n とする)．ここで，z_n は学習時のみ各サンプルに対して与えられ，予測時には与えられない．すなわち，z_n は補助情報，とくに特権情報 [1] である．

評価指標としては，誤判別コストのクラス間の違いを考慮した重み付き精度 (WA: weighted accuracy) を用いる．

$$\mathrm{WA}(\{(z_n, f(x_n))\}) = \frac{1}{N} \sum_n C_+ I_{z_n \geq 0} I_{f(x_n) \geq 0.5} + C_- I_{z_n < 0} I_{f(x_n) < 0.5}, \quad (1)$$

ここで，N はサンプル数，C_+, C_- はそれぞれ正例と負例に対する誤判別コストであり，問題ごとに設定される定数である．とくに実験では，これらを頻度の逆数に比例する値に設定した場合，すなわち $C_+ = N/2N_+, C_- = N/2N_-$ の場合について

評価する．

$$\mathrm{BA}(\{(z_n, f(x_n))\}) = \frac{1}{N}\sum_n \frac{N}{2N_+} I_{z_n \geq 0} I_{f(x_n)\geq 0.5} + \frac{N}{2N_-} I_{z_n<0} I_{f(x_n)<0.5}, \quad (2)$$

ここで $N_+ = \sum_n I_{z\geq 0}, N_- = \sum_n I_{z<0}$ はそれぞれ正例数，負例数である．これを均衡化精度 (BA: balanced accuracy) と呼ぶ．$1-\mathrm{BA}$ は均衡化誤差率 (BER: balanced error rate) と呼ばれ，しばしば不均衡判別問題において採用される [2]．

3 関連研究

我々の問題設定は新規のものであるが，いくつかの関連する研究を以下に挙げ，違いを明確化する．

3.1 確率ラベルによる教示に基づく学習

まず，二値よりも質の高いラベルが与えられるという点では，確率ラベル (ソフトラベル)$\{s_n = p(y=1|x_n)\}$ が与えられる設定はよく研究されている [3, 4, 5]．この設定では，確率ラベル s に対するガウス過程回帰の適用 [5] や，サンプルペア間の確率ラベルの大小関係を用いたランク学習 [4] などにより学習のサンプル効率化が可能であることが示されている．確率ラベルはたとえば，クラウドソーシングにより集められたラベルを平均することにより与えられることが想定されている．我々の設定における正例度合い z は確率ラベル $p(y=1|x)$ と強い関係があると考えられるが，それそのものではない．そのためたとえば直接に z を回帰し，その予測値に閾値を適用する，すなわち $\hat{y} := I(\hat{z} \geq \theta')$ とすることで高精度な判別が可能となるかどうかはサンプル数や分布に依存し，性能は保証されない．

3.2 特権情報を用いた学習

二値のラベルに加えて学習時のみ各サンプルに与えられる情報であって，ラベルに強い相関を持つことが想定されるものは特権情報 (PI: privileged information) と呼ばれる．特権情報を用いた学習 (LUPI: learning using privileged information) は，はじめ SVM においてスラック変数を推定する目的で導入され [1, 6]，その後，一般化蒸留 (GD: generalized distillation)[7] により一般のモデルの学習に拡張された．GD では，まず “教師モデル” をラベルと PI $\{y_n, x_n^*\}_n$ から学習し，その後 “生徒モデル” f を以下の損失関数の最小化により学習する．

$$L_{S,T}^{\mathrm{GD}}(f) = \frac{1}{N}\sum_n \sigma(g_t(x^*)/T)\log f(x_n) + \sigma(-g_t(x^*)/T)\log(1-f(x_n)), \quad (3)$$

ここで，g_t は教師モデルの決定関数，T は温度と呼ばれるハイパーパラメタ，$\sigma(a) := 1/(1+\exp(-a))$ はシグモイド関数である．

第 4章にて説明する提案手法は，正例度合い z を PI とした GD と，クラス均衡化の組み合わせである．ただし，GD を含め LUPI は主に学習率の向上，すなわちサンプル数を増やしたときに期待損失が最適なモデルパラメタに対する期待損失へ収束する速さを問題としているのに対し，我々の目的は正例が少ない状況での精度向上である．

3.3 コスト考慮型学習

PI を用いない通常の設定における不均衡データの判別においてよく知られた従来手法の一つはコスト考慮型学習 [8] である．その損失関数はクラスごとの誤判別コスト C_+, C_- によって各サンプルを重みづけすることによって得られる．

$$L_S(f) = \frac{1}{N}\sum_n C_+ y\log f(x_n) + C_-(1-y)\log(1-f(x_n)). \quad (4)$$

コスト考慮型学習の損失関数はほぼ $1-\mathrm{WA}$ の凸緩和となっていることから，サンプル数が十分な状況では良い WA を達成する．しかし，正例数 N_+ が小さく正例の重み C_+ が大きいとき推定分散が大きくなる．第 4章では，GD とコスト考慮型学習の組み合わせによる提案手法の損失を導入し，その最小化における期待余剰損失の上界を与え，従来のコスト考慮型学習に対する優位性を議論する．第 5章では，実データを用いた実験によって従来のコスト考慮型学習と提案手法を様々な正例比率で比較し，正負例の偏りが非常に大きい場合における提案手法の優位性を実験的に示す．

4 提案手法

提案する損失関数は，GD 損失 (3) とコスト考慮型損失 (4) の組み合わせ，また GD の導入による正負例の期待重みの変化を再均衡化することにより，下記の代理損失として与えられる．

$$L_{S,T}(f) = \frac{1}{N}\sum_n C_{T,+}\sigma(z_n/T)\log f(x_n) + C_{T,-}\sigma(-z_n/T)\log(1-f(x_n)), \quad (5)$$

ここで $C_{T,+}, C_{T,-}$ はそれぞれ再均衡化されたコストパラメタで，下記で与えられる．

$$C_{T,+} = C_+\frac{p_+}{p_{T,+}}, \quad C_{T,-} = C_-\frac{p_-}{p_{T,-}}, \quad (6)$$

ここで p_+, p_- はそれぞれ正例比率，負例比率，$p_{T,+}, p_{T,-}$ はそれぞれ GD によるソフトラベルの有効正例比率，有効負例比率であり，$p_+ := \mathbb{E}[y], p_{T,+} := \mathbb{E}[\sigma(z_n/T)]$ のように与えられる．

上記の損失関数におけるソフトラベル $\sigma(z_n/T)$ は，パラメタ T を 0 に漸近させたとき元のラベル y に一致する，すなわち，$\sigma(z/T) \to I(z\geq 0) = y$ as $T \to 0$．よって，提案手法の損失関数はコスト考慮型学習の損失関数を $T\to 0$ の特殊ケースとして含む．

この損失関数の最小化に対する期待余剰損失は以下のように上から抑えられる．ただし，本解析においてはモデルクラスを有界線形クラスとする．

定理 4.1 (提案損失関数に対する期待余剰損失上界)．w_T^* を最適なパラメタ，すなわち期待代理損失を最小化するパラメタとし，$\hat{w} : \|\hat{w}\|_2 \leq B$ を経験代理損失 *(5)* を最小化するパラメタとする．また，説明変数 x は確率 1 で有界とする，すなわち，$p(\|x\|_2 \leq X) = 1$．このとき，期待余剰損失は以下のように上から抑えられる．

$$\mathop{\mathbb{E}}_S[L_T(\hat{w}_T) - L_T(w_T^*)] \leq \frac{2BX}{\sqrt{N}}\sqrt{C_+^2\frac{p_+^2}{p_{T,+}} + C_-^2\frac{p_-^2}{p_{T,-}}},$$

ここで，$\mathbb{E}_S$ は学習用サンプルのとり方に関する期待値とする．

証明は紙面の都合上割愛する．上記に対し，均衡化精度を最大化するコストパラメタ設定，すなわち

$$C_+ = \frac{1}{2p_+} \text{ and } \quad C_- = \frac{1}{2p_-}, \tag{7}$$

とすると以下を得る．

系 4.1.1 (均衡化コスト設定における期待余剰損失上界)．

$$\mathop{\mathbb{E}}_{S}[L_T(\hat{w}) - L_T(w_T^*)] \leq \frac{BX}{\sqrt{N}}\sqrt{\frac{1}{p_{T,+}} + \frac{1}{p_{T,-}}}. \tag{8}$$

上記定理および系から次のことがわかる．不均衡な設定のとき，すなわち正例が非常に少なく，かつ正例に対するコストが大きく設定されているとき，従来のコスト考慮型学習では正例比率 $\lim_{T\to 0} p_{T,+} = p_+$ が小さいことから上界が大きくなる．一方で，z の情報を用いてニアミス正例，すなわち z の値が比較的大きい負例を部分的に正例として扱うことによって有効正例比率 $p_{T,+}$ を増加できた場合，上記上界は改善される．

第 5章の実験結果からも，正負例比率が大きく偏っている場合に精度が改善されることが示される．

5 実験

第 4章で考察した提案手法の利点を実データを用いた実験により検証する．提案手法は実数値の正例度合い z が得られることを仮定しているため，実数値の目的変数を持つ回帰用データセットであり，かつ高度に不均衡な設定を含む実験を行うため，サンプル数の大きいデータセットとして GPU カーネルパフォーマンスデータセット [9, 10] を用いた．本データセットは様々な GPU のカーネルパラメタで演算を実行したときの経過時間が，各パラメタの組み合わせに対して 4 回ずつ測定されている．我々はこれを，高速な演算を実現するパラメタの組み合わせを見つける問題とした．すなわち，測定された経過時間 $\{\tau_1, \ldots, \tau_4\}$ から計算した平均実行速度 $z := \frac{4}{\sum_i \tau_i}$ を正例度合いとし，その上位 $100p_+\%$ が正例となるようそれぞれ閾値 θ を設定した．パラメタの種類，すなわち説明変数の次元は 14，平均実行速度に変換後のサンプル数は 60,400 である．正例度合い z の分布を図 2に示す．右裾部では概ね右肩下がりの分布形状となっている．このような分布であれば，右裾部に閾値 θ を設定したとき，ニアミス正例（負例であって正例度合いが高いもの）はニアミス負例よりも多くなる，すなわち適度なパラメタ T のもとで有効正例比率 $p_{T,+}$ は本来の正例比率 p_+ と比較して増加すると考えられる．

モデルには，提案手法，従来のコスト考慮型学習ともに L2 正則化項付きロジスティック回帰を用いた．正則化の強さと提案手法のハイパーパラメタ T の選択，および評価には入れ子式交差検証 [11] を用いた．その際，外側ループは 5 分割，内側ループは 2 分割とした．すなわち，データセットを 5 分割し，1 つを検証用，4 つを学習用とし，学習用データでさらに 2 分割交差検証によってハイパーパラメタを選択，その後選択したハイパーパラメタで学習用データ全体を用いて再度学習し，検証用データで検証を行う手続きを 5 回行った．さらに，分割方法を変えて 4 回ずつ前記操作を行い，計 20 回の検証を行った．

結果を図 3に示す．正例比率 p_+ が比較的大きい領域では提案手法は既存手法と同等の精度（BA）であるのに対し，正例比率が小さい領域では既存手法は精度が悪化する一方，提案手法では精度悪化が抑えられている．これは，提案手法では正例度合いの情報に基づいてニアミス正例を部分的に正例と扱うことにより有効正例比率 $p_{T,+}$ が高まり，期待余剰損失 (8) が軽減されたためと考えられる．

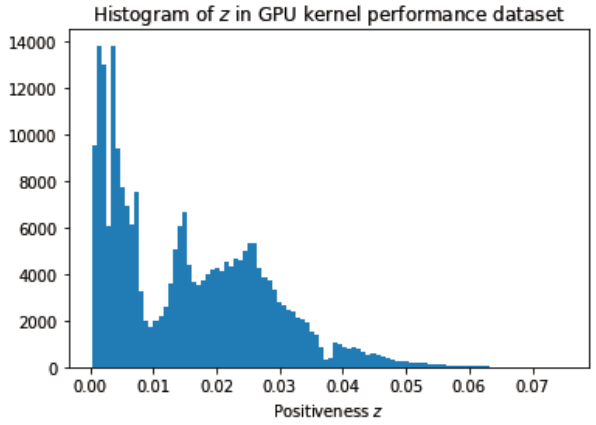

図 2: GPU カーネルパフォーマンスデータセットにおける正例度合い z のヒストグラム．

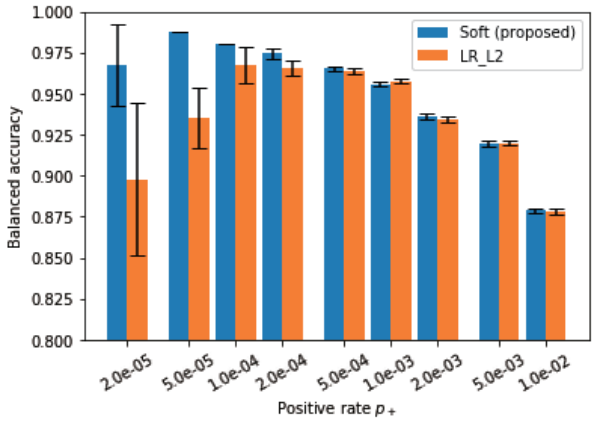

図 3: 提案手法と従来のコスト考慮型学習の均衡化精度の比較．高度に不均衡なとき（正例比率 $p_+ := \sum I(z \geq 0)/N$ が小さいとき）提案手法が高精度となっている．誤差バーは標準誤差を表す．

6 結論

我々は，実数値目的変数が与えられるが最終目的は判別であるという問題設定を新たに導入し，コスト考慮型学習と一般化蒸留の組み合わせによる損失関数を提案した．提案手法は既存のコスト考慮型学習と比べ，正負例比率が非常に偏っている場合でも均衡化精度の点で高性能となることを理論及び実験の両面から検証した．本手法は正例度合いが取得できるという強い仮定に依存している．一方で本結果は，災害や事故など正例が稀な現象に対応する場合でも，負例の中でも危険であった事例とその度合いを取得することができれば高精度な判別が可能となることを示唆し，そのようなデータの取得を推奨する．

参考文献

[1] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

[2] Xue-wen Chen and Michael Wasikowski. Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 124–132. ACM, 2008.

[3] Quang Nguyen, Hamed Valizadegan, Amy Seybert, and Milos Hauskrecht. Sample-efficient learning with auxiliary class-label information. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1004. American Medical Informatics Association, 2011.

[4] Yanbing Xue and Milos Hauskrecht. Learning of classification models from noisy soft-labels. In *ECAI*, pages 1618–1619, 2016.

[5] Peng Peng, Raymond Chi-Wing Wong, and Phillp S Yu. Learning on probabilistic labels. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 307–315. SIAM, 2014.

[6] Vladimir Vapnik, Rauf Izmailov, Alex Gammerman, and Vladimir Vovk. Learning Using Privileged Information: Similarity Control and Knowledge Transfer. *Journal of Machine Learning Research memory of Alexey Chervonenkis*, 16:2023–2049, 2015.

[7] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. UNIFYING DISTILLATION AND PRIVILEGED INFORMATION. In *Proceedings of the International Conference on Learning Representations*, pages 1–10, 2016.

[8] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[9] Cedric Nugteren and Valeriu Codreanu. Cltune: A generic auto-tuner for opencl kernels. In *Embedded Multicore/Many-core Systems-on-Chip (MCSoC), 2015 IEEE 9th International Symposium on*, pages 195–202. IEEE, 2015.

[10] Rafael Ballester-Ripoll, Enrique G Paredes, and Renato Pajarola. Sobol tensor trains for global sensitivity analysis. *arXiv preprint arXiv:1712.00233*, 2017.

[11] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91, 2006.