

不均衡データセットを用いた回帰問題における 損失関数の検討

吉川 寛樹^{1,a)} 内山 彰^{1,b)} 東野 輝夫^{1,c)}

概要: 機械学習において連続値を推定する回帰問題は広く用いられており、科学やヘルスケアなど様々な分野において応用が盛んである。しかしながら実環境において収集した訓練データの分布には、しばしば偏りが発生し、特に異常時のデータは平常時のデータと比較してサンプル数が少なくなる。このような正解ラベルの分布の偏りは、機械学習における推定器の性能の低下を招く。本研究では正解ラベルの分布が偏った不均衡なデータセットに対し、機械学習を行う際の推定値の偏りを軽減するための損失関数を提案する。提案手法では正解ラベルの連続的な分布をカーネル密度推定により推定することで、データの希少性を表す関数を算出し、その関数を用いて推定誤差に重みを付けた損失関数を用いる。性能評価では、連続値を正解ラベルとした8種の公開不均衡データセットに対し提案手法を適用することによる推定モデルの性能の向上を確認した。

1. 研究背景

機械学習は情報科学や統計学だけでなく非常に広範な分野において研究、応用されている。近年では深層学習を用いて複雑な処理を伴う手法が多く開発されているが、その処理の内容がパッケージ化されることにより、専門的な知識を持たずとも利用が可能となり広く普及している。しかしながら、訓練のされ方によらず推定器は何らかの値を出力するため、利用者は推定器がもっともらしい値を出力していれば正しく動作していると思いついてしまう。そのような理由から、利用者は推定器が意図した通りに訓練されていなくても、それに気づくことが難しい。気づくことが難しい問題の一つに、不均衡データの訓練に起因する推定値の偏りがある [1]。不均衡データと呼ばれる正解ラベルの分布に偏りがあるデータを訓練に用いると、推定器は多数派のデータを必要以上に推定値として出力しやすくなる。その結果、本来少数派のデータに対しても多数派として推定してしまう推定器が訓練される。

そのような問題を解決するために、不均衡データを訓練に用いるための手法が、特に分類問題においては多く研究されている。手法は2種類に大別される。これらの手法の概要を図1に示す。1つ目はデータバランシングである。この手法は機械学習の前処理として、サンプル数を増減さ

せることでクラス間のサンプル数の差を小さくし、不均衡を緩和する手法である。2つ目は訓練時のペナルティにサンプル数に応じた重みを付ける手法である。この手法は少数派クラスと多数派クラスのそれぞれのサンプル数に応じて、少数派クラスの訓練時のペナルティを相対的に大きくする手法である。これらの手法によりサンプル数の違いによってクラス間の推定精度に差が出ることを抑制する。例えば、2クラス間の分類問題において不均衡データを扱う際には、一方が多数派クラス、もう一方が少数派クラスとなるため、前述の手法を用いることで相対的に多数派クラスを軽視し、少数派クラスを重視するような訓練を行う。

一方で、機械学習においては回帰問題も非常に広範な分野において研究、応用されている。回帰問題は訓練データの正解ラベルを連続値として扱い推定器を構築する。応用事例の多くでは、訓練データは実世界の観測に基づいて収集されるため、有限個の観測から得られた離散的な正解ラベルを用いて連続値を推定できるよう推定器を訓練する必要がある。このようなデータにおいて正解ラベルは離散的な分布となり、少なからず不均衡が発生する。この不均衡が特に顕著なデータセットを用いて訓練を行うと、分布が集中している近辺の値を出力しやすい推定器となってしまう。サンプル数の偏りが大きいほどこの傾向が強くなり、分布の少ない値は出力されにくくなる。しかしながら異常検知など分布の少ない値の推定が重要となる場合は、分布の少ない値に対し高精度な推定が要求される。本研究ではこのようなデータセットの不均衡に起因する推定器の精度

¹ 大阪大学大学院情報科学研究科

^{a)} h-yoshikawa@ist.osaka-u.ac.jp

^{b)} uchiyama@ist.osaka-u.ac.jp

^{c)} higashino@ist.osaka-u.ac.jp

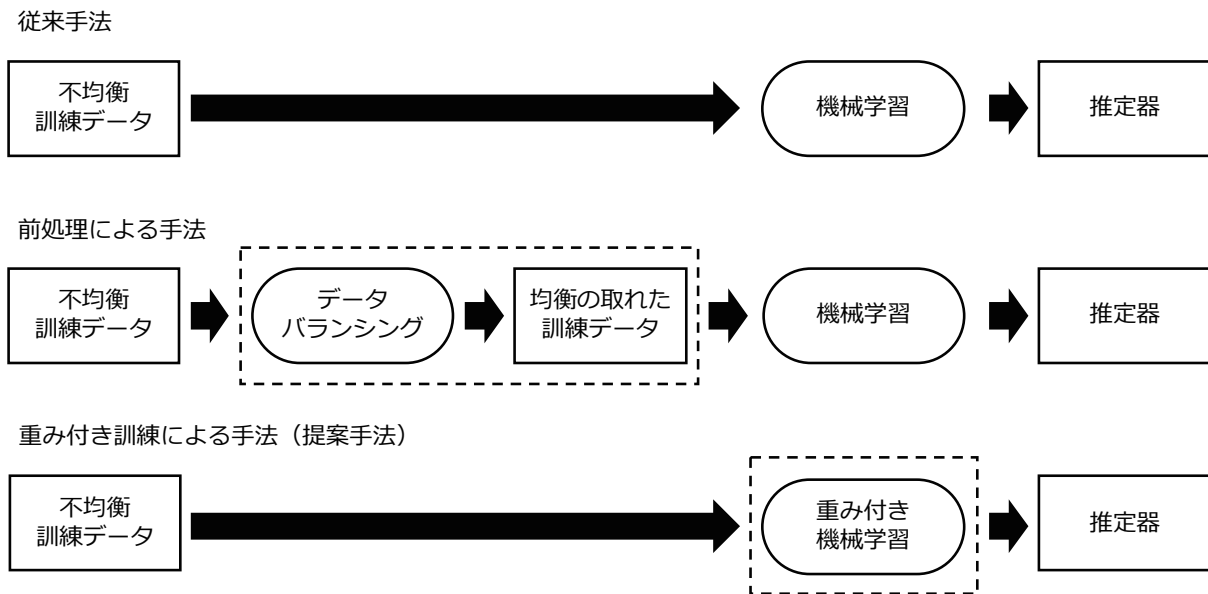


図 1 不均衡データへの対策手法の概要

の低下を防ぐための手法の構築を目的とする。

回帰問題においても前述の2種類の不均衡データへの対策手法が研究されている。1つ目のデータバランシングによる手法では、回帰問題として扱う連続的なデータをしきい値により少数派クラスと多数派クラスに分割することで、クラス分類におけるデータバランシングであるSMOTE [2]を応用する手法が提案されている。こちらの手法はデータバランシングの特徴である、前処理として組み込むために訓練の方法に依らずシステムに組み込むことが可能である。しかし、手法の多くは横断面データと呼ばれる、ある時点でのセンサによる測定値のような、瞬間を切り取ったデータに対して用いることが前提とされており、時系列のような高次元のデータへの対応は想定されていない。また、2つ目の訓練時のペナルティに重みを付与する手法も提案されている。このような訓練時のペナルティに重みを付与する手法は、データの生成を伴わないため時系列や画像等の高次元データに対しても使用可能であるが、パラメータの調整に専門知識を必要とすることや、使用できる正解ラベルの分布が限られているために実用的に広く普及するには至っていない。

本研究では、ユーザがパラメータを設定することなく訓練時のペナルティに付与する重みを算出することで、機械学習に関する専門知識を必要とせず、特徴量データの種類によらず使用可能な、回帰問題における不均衡データセットへの対応手法を提案する。提案手法は、平均絶対誤差や平均二乗誤差等、回帰問題で一般的に用いられる訓練時のペナルティに対し、データの希少性に基づいて重みを付与する。データの希少性は訓練データに含まれる正解ラベルの分布に基づいて算出する。性能評価では、データの密度の影響を受けず推定性能を評価することが可能な指標を提

案し、平均絶対誤差が減少することを確認した。

2. 関連研究

2.1 分類問題における不均衡データへの対策手法

分類問題においては不均衡データへの対策手法が数多く提案されている。データバランシングの基本的なアイデアはデータの前処理によりそのサイズと多様性を増やすことである [3]。実在するデータに対して内挿的に新たにデータを生成する手法であるSMOTE [2]を応用した手法が数多く存在し、Adaptive Synthetic sampling approach (ADASYN) [4] や Borderline-SMOTE [5] などが提案されている。SMOTEに基づく拡張手法は、SMOTEの補間手順を、クラスターリングや確率関数などの他のより複雑なものに置き換えることで推定精度を向上させる手法が多く存在する。さらにSMOTEの後処理として、SMOTE + Tomek [6] や SMOTE + ENN [7] などが提案されている。これらの手法はクラス間の境界を明確にするために、データを増やした後にデータセットから不要なデータを削除する。これらはSMOTEなどのオーバーサンプリング手法に対して、アンダーサンプリング手法の一種である。

また、近年では生成的モデル [8] と呼ばれる深層学習に基づいたオーバーサンプリング手法も提案されている。これらの手法はGenerative Adversarial Networks (GANs) [9] や Variational Autoencoder (VAEs) [10] などの深層学習を用いてデータを生成する。これらのアルゴリズムは、実データの分布を学習することでガウス分布に従うノイズからデータを生成する関数をモデル化する。これらのアプローチは、コンピュータビジョンやサイバーセキュリティなどのいくつかの分野で、現実に即した値を生成することが示されている [11]。一方で、生成的モデルは学習が必要

である。つまり不均衡なデータセットを生成的モデルの学習に用いると、一般的な値に対して生成的モデルが過剰に適合し、一般的な値を生成しやすいモデルが訓練される可能性がある。

また、訓練時のペナルティに重みを与える手法も存在する。Zadrozny ら [12] は誤分類時のコストに応じてデータの分布に重みを与える手法を提案している。これにより分類アルゴリズムに重みの情報を与えて訓練を行うことができる。以上のような分類問題における不均衡データへの対策手法は、連続値を推定する回帰問題においては用いることが難しい。これは正解ラベルが離散的な場合にはデータ数に応じて重みを与えることができるが、回帰問題では正解ラベルが連続的な分布を持ち、同様の手法で正解ラベルに重みを与えることが難しいためである。

2.2 回帰問題における不均衡データへの対策手法

回帰問題においても分類問題と同様に、前処理による手法と訓練時のペナルティに重みを与える手法の2種類の不均衡データへの対策手法が研究されている。1つ目のデータバランシングによる手法では、回帰問題として扱う連続的なデータをしきい値により少数派クラスと多数派クラスに分割する手法が提案されている。このように分割することで、クラス分類におけるデータバランシングであるSMOTE [2] を応用する手法であるSMOTER [13] が提案されている。さらに、SMOTER に対し、実在するサンプルにランダムノイズを加えることでデータを生成する手法を組み合わせたSMOIGN [14] も提案されている。これらのようにシステムへの組み込みやすさと性能の両面で優れているSMOTEを応用した手法が実用的には多く用いられている。

SMOTEはデータを増やす際に内挿的にサンプルを生成する手法であるため、1つのサンプルに偏った訓練がされにくい。加えて、データバランシングの特徴である、前処理として組み込むために訓練の方法に依らずシステムに組み込むことが可能である。そのためSMOTEのように元々のデータの傾向を捉えて、似た傾向を持つサンプルを生成する手法が近年では数多く提案されている。そのような手法の多くは横断面データと呼ばれる、ある時点でのセンサによる測定値のような、瞬間を切り取ったデータに対して用いることが前提とされており、時系列のような高次元のデータは想定されていない。さらに、これらの手法は、連続値である正解ラベルをユーザが与えるしきい値により分割することでクラス分類の問題に帰着させているため、少数派クラスに振り分けられたサンプルに傾向が似たサンプルを生成することから、少数派クラスの中で多く分布するサンプルの傾向を反映したサンプルを生成しやすくなってしまふ。その結果、最初に分割した多数派クラスと少数派クラスの中でも、手法適用後に再び不均衡が発生すること

となり、不均衡の緩和には一定の効果があるが根本的な解決とはならない。

そのため、特徴量を抽出する対象が時系列や画像など高次元データである場合には用いることが困難である。また、個々のサンプルに対するオーバーサンプリングのされやすさとアンダーサンプリングのされやすさの確率をデータの密度に応じて変化させることで分布の均衡を取る手法が存在する [15]。この手法では、しきい値による多数派データと少数派データへの分割を行わず正解ラベルを連続値として扱うことでデータの大局的な均衡を取る。その一方で、連続的な値に対して確率的にサンプルを生成または除去するという性質から、データバランシング後の分布にも少なからず不均衡は発生するため不均衡の根本的な解決とはならない。さらにこれらのデータバランシング手法は機械学習の前処理として導入されるため、機械学習の手法に依らず不均衡データに起因する問題を緩和できる一方で、オーバーサンプリングやアンダーサンプリングの比率等のユーザによるパラメータ設定が必要であるためユーザが専門知識無しに導入することは容易ではない。

これらの手法に対し、回帰問題において訓練時のペナルティに重みを付与する手法も存在する。Utility-based regression [16] では、データの分布に基づいて、サンプルを学習する際のbenefitとcostを反映したutilityと呼ばれる値を最大化することで推定器を訓練する。しかしながら、benefitとcostの重みの比率をパラメータとして設定する必要があり、専門知識に基づいてユーザの要求に推定器をフィッティングさせることを必要とする。さらにデータの希少性を表現するrelevance functionを生成する際にもデータの分布を仮定しており、多様な分布のデータへ対応するためにはユーザ自身でrelevance functionを与える必要がある。このような訓練時のペナルティに重みを付与する手法は、データの生成を伴わないため時系列や画像等の高次元データに対しても使用可能であるが、パラメータの調整に専門知識を必要とすることや、使用できる正解ラベルの分布が限られているために実用的に広く普及するには至っていない。本研究では、訓練時のペナルティに付与する重みをパラメータに依存しない手法により算出することで、正解ラベルの分布や特徴量データの種類の依らず使用可能な、回帰問題における不均衡データセットへの対応手法を提案する。

3. 提案手法

本研究では、訓練時のペナルティに付与する重みを算出することで、特徴量データの種類の依らず使用可能な回帰問題における不均衡データセットへの対応手法を提案する。提案手法は、正解ラベルの分布を仮定しないことで機械学習の専門知識を持たないユーザもパラメータ調整を必要としないノンパラメトリックな手法である。これは確率密度

Algorithm 1 Relevance function generation.

Input: y - Actual label list
Output: f_r - Relevance function

```

# Kernel density estimation with Scott's rule
 $f_{dens} \leftarrow \text{KERNELDENSITYESTIMATION}(y)$ 
# Normalize density function
 $f_{norm} \leftarrow f_{dens} / \max(f_{dens})$ 
# Generate relevance function as complementary standardized density function
 $f_r \leftarrow 1 - f_{norm}$ 
return  $f_r$ 
    
```

関数の推定手法であるカーネル密度推定 (KDE) [17], [18] に、実世界のデータに対する適合性が経験的に保証されている Scott のルールを用いて正規分布の形を仮定することで、パラメータを必要とせずに正解ラベルの確率密度関数を推定するというアイデアに基づいている。提案手法は、平均絶対誤差や平均二乗誤差等、回帰問題で一般的に用いられる訓練時のペナルティに対し、データの希少性に基づいて重みを付与する。

データの希少性は訓練データに含まれる正解ラベルの分布に基づいて、分布の形を仮定しない手法である KDE を用いて、下記の定義に基づき確率密度関数 f_{dens} を算出する。

$$f_{dens}(x) = \frac{1}{nh} \sum_{i=0}^n K\left(\frac{x-x_i}{h}\right) \quad (1)$$

ここで n はデータ数、 h はバンド幅、 K はカーネル関数である。さらに提案手法では、カーネル関数として一般的に用いられるガウス関数

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{(-x^2/2)} \quad (2)$$

を用い、バンド幅 h は、経験的に算出する手法である Scott のルール [19]

$$h \approx 1.06un^{(-1/5)} \quad (3)$$

を用いることで、パラメータ設定を必要とすることなく KDE を実行する。ここで u は訓練データの正解ラベルの不偏標準偏差である。以上の手順により取得した正解ラベルの連続的な分布に基づいて正解ラベルの入力に対しその希少性を出力する relevance function [16] を生成する。提案手法における relevance function f_r の生成手法を Algorithm 1 に示す。KDE により取得した確率密度関数 f_{dens} を標準化することで、最大値 1、最小値 0 となった関数 f_{norm} を生成する。関数 f_{norm} を 1 から減算することで、少数派データほど 1 に近く、多数派データほど 0 に近い値を返す relevance function f_r を生成する。

以上の手順により生成した relevance function f_r を用いて、機械学習における重み付き損失関数 $RExE$ [20] を以下のように定義する。

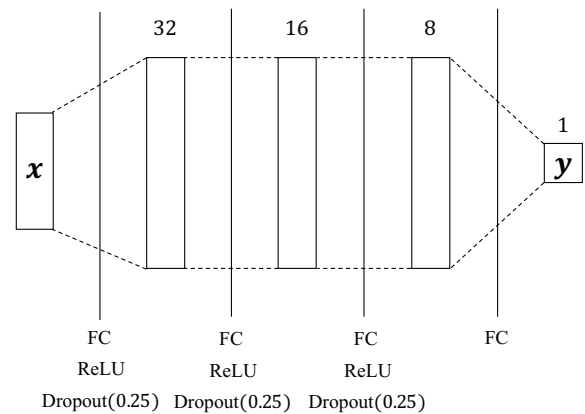


図 2 ニューラルネットワークによる推定器

$$RExE = \frac{1}{n} \sum_{i=0}^n f_r(y_i) L(y_i, \hat{y}_i) \quad (4)$$

ここで、 n はデータ数であり、 y_i, \hat{y}_i はそれぞれ正解ラベルとそのサンプルに対応する推定値である。また、 $L(y_i, \hat{y}_i)$ は、任意の誤差関数であり回帰問題では二乗誤差や絶対誤差が一般的には用いられる。これにより、重み付き損失関数 $RExE$ を最小化することで、少数派データに対しても誤差の小さい推定が可能である推定器を訓練する。

4. 性能評価

4.1 評価環境

本研究では提案手法を評価するために連続値を正解ラベルとした複数のデータセットを用いて評価を行う。横断面データによる連続値の推定を行うために、文献 [15] で使用されている公開データセットを用いる。本研究では図 2 に示す、ニューラルネットワークによる推定器を使用するため、欠損値を除去した後にサンプル数が 500 以上である 8 種のデータセット*1(Abalone, boston, fuelCons, heat, availPwr, cpuSm, bank8FM, Accel) を評価に用いる。

評価指標には、一般的に回帰問題の評価指標として用いられる平均絶対誤差 (MAE) に加え、データの分布密度によらず推定値の正確さを評価するために、MAE に relevance function による重みを乗算した重み付き平均絶対誤差 (MWAE) を用いる。MWAE の定義を以下に示す。

$$MWAE = \frac{1}{n} \sum_{i=0}^n f_r(y_i) |y_i - \hat{y}_i| \quad (5)$$

また比較手法として損失関数に平均二乗誤差を用いる手法をベースラインとする。

4.2 評価結果

それぞれのデータセットに含まれる正解ラベルのヒストグラムと relevance function を図 3 に示す。各データセットは多数派データと少数派データが存在する不均衡データ

*1 <https://paobranco.github.io/DataSets-IR/>

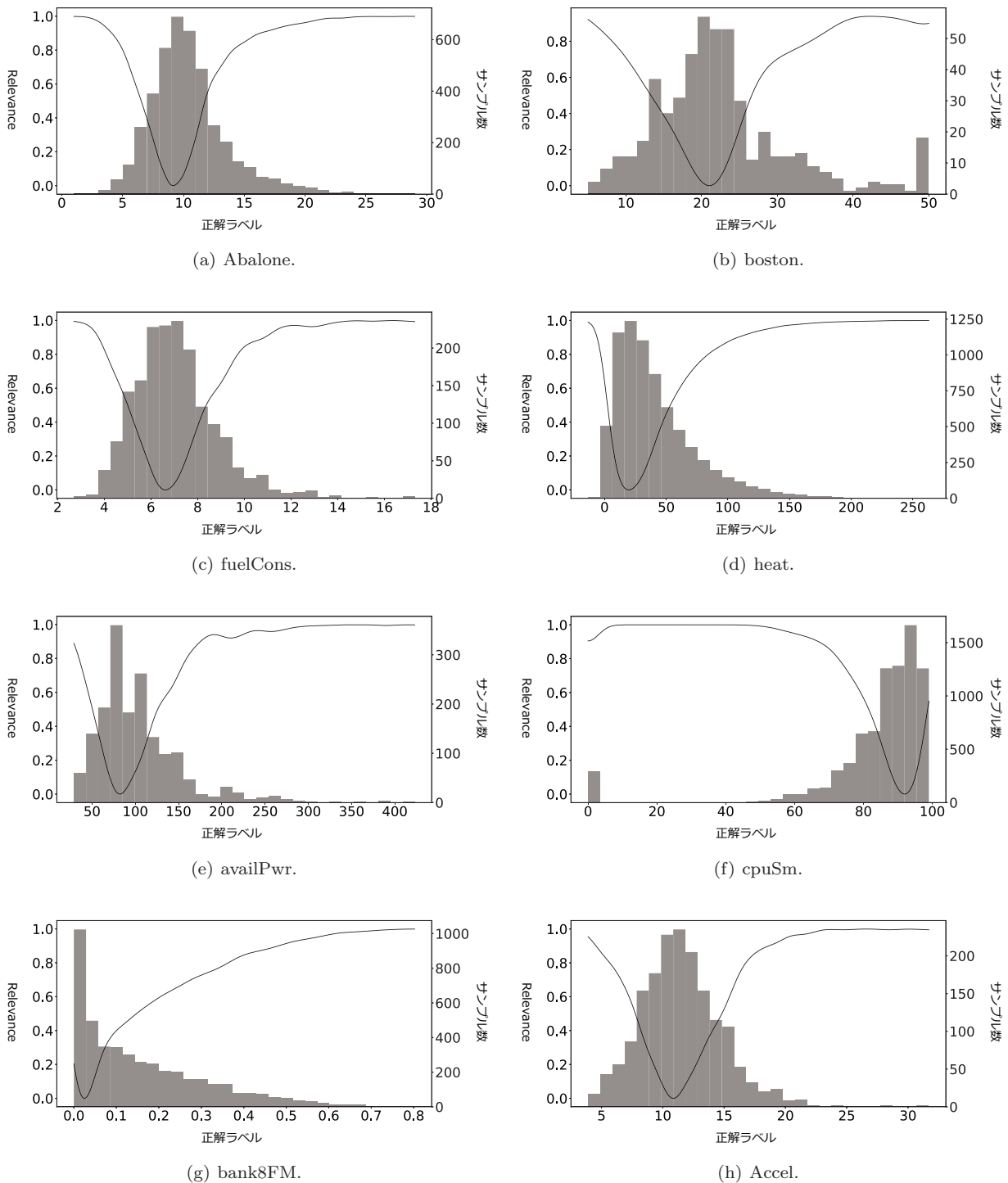


図 3 各データセットに含まれる正解ラベルのヒストグラムと relevance function. 各グラフは左軸が relevance function, 右軸はヒストグラムのサンプル数, 横軸は正解ラベルの値を示す.

であり, 3節で提案した relevance function を正解ラベルの分布に基づき生成できていることが分かる. 各データセットに対して生成した relevance function により, それぞれのデータセットに対する推定を行った際の各評価指標の値を表 1 に示す. また, 回帰問題における損失関数に一般的に用いられる MSE を用いた手法をベースライン手法とし

て比較する.

MAE については提案手法によって減少したデータセットは 3 種に留まっているが, 大きく誤差が改善されたデータセットも存在する. 一方で MWAE においては 8 種のデータセットの内 7 種のデータセットにおいて改善を確認した. また, 改善が確認できなかったデータセットに対し

表 1 各データセットを用いた推定結果. 上部が提案手法, 下部がベースライン手法.

Dataset	MAE	MWAE
Abalone	2.23	0.76
	1.66	0.82
boston	4.14	1.66
	3.42	1.83
fuelCons	1.95	0.72
	3.15	1.18
heat	11.32	3.85
	10.33	5.10
availPwr	17.59	6.11
	33.47	15.55
cpuSm	5.94	1.683
	4.45	1.679
bank8FM	0.046	0.021
	0.043	0.023
Accel	1.43	0.48
	3.41	1.31

ても MAE の増加は 0.2% に留まっており, 大きな誤差の増加は確認されなかった.

さらに, 少数派データに対する推定誤差を詳細に調査するため, relevance function があるしきい値以上となる正解ラベルを持つデータに対する MAE を算出しプロットした結果を図 4 に示す. いずれのデータセットに対しても, relevance function のしきい値の上昇に伴い MAE も増加する傾向が見られるが, 提案手法は MAE の増加が比較的緩やかであり, 少数派データほど MAE がベースライン手法に比べ小さくなる. 特に relevance function のしきい値が 0.3 以上のときには全てのデータセットで, 提案手法の MAE が小さい結果となった. ただし relevance function のしきい値が 0.9 以上のときは, データ数が非常に少ないことから MAE のばらつきが大きいので, ベースライン手法が提案手法を上回るデータセットも存在した.

5. 結論

本研究では, 機械学習において連続値を推定する回帰問題における, 不均衡データを訓練に用いるための損失関数の検討を行った. 提案手法では, 多数派データに比べ少数派データに重みを与えることで, 正解ラベルの分布が不均衡であることによる訓練の偏りを緩和する. 性能評価により, 8 種の公開データセットに対して損失関数に重みを与えない従来手法と提案手法を比較した結果, 提案手法は少数派データに対する平均絶対誤差が小さく, 特にデータの希少性を表す relevance function のしきい値が 0.3 以上のデータに対しては全てのデータセットで, 提案手法の MAE が小さい結果となった. 今後は多数派データについても誤差の小さい推定を行うため, relevance function の値に基づいたモデルの切り替えを行うことを考えている.

参考文献

- [1] Paula Branco, Luís Torgo and Rita P. Ribeiro: A Survey of Predictive Modeling on Imbalanced Domains, *ACM Comput. Surv.*, Vol. 49, No. 2 (online), DOI: 10.1145/2907070 (2016).
- [2] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer: SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357 (online), DOI: 10.1613/jair.953 (2002).
- [3] Sergey I. Nikolenko: Synthetic Data for Deep Learning (2019).
- [4] Haibo He, Yang Bai, Garcia, E. A. and Shutao Li: ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328 (online), DOI: 10.1109/IJCNN.2008.4633969 (2008).
- [5] Hui Han, Wen-Yuan Wang and Bing-Huan Mao: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, Vol. 3644, pp. 878–887 (online), DOI: 10.1007/11538059_91 (2005).
- [6] Gustavo Batista, Ana Bazzan and Maria-Carolina Monard: Balancing Training Data for Automated Annotation of Keywords: a Case Study., pp. 10–18 (2003).
- [7] Gustavo E. A. P. A. Batista, Ronaldo C. Prati and Maria Carolina Monard: A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data, *SIGKDD Explor. Newsl.*, Vol. 6, No. 1, pp. 20–29 (online), DOI: 10.1145/1007730.1007735 (2004).
- [8] Harshvardhan GM, Mahendra Kumar Gourisaria, Manjusha Pandey and Siddharth Swarup Rautaray: A comprehensive survey and analysis of generative models in machine learning, *Computer Science Review*, Vol. 38, p. 100285 (online), DOI: 10.1016/j.cosrev.2020.100285 (2020).
- [9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio: Generative Adversarial Nets, *Advances in Neural Information Processing Systems*, Vol. 27, Curran Associates, Inc., pp. 2672–2680 (2014).
- [10] Diederik P. Kingma and Max Welling: Auto-Encoding Variational Bayes, *Proceedings of 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada* (2014).
- [11] Md Zahangir Alom, Tarek M. Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Mahmudul Hasan, Brian C. Van Essen, Abdul A. S. Awwal and Vijayan K. Asari: A State-of-the-Art Survey on Deep Learning Theory and Architectures, *Electronics*, Vol. 8, No. 3 (online), DOI: 10.3390/electronics8030292 (2019).
- [12] Bianca Zadrozny, John Langford and Naoki Abe: Cost-sensitive learning by cost-proportionate example weighting, *Third IEEE International Conference on Data Mining*, pp. 435–442 (online), DOI: 10.1109/ICDM.2003.1250950 (2003).
- [13] Luís Torgo, Rita Ribeiro, Bernhard Pfahringer and Paula Branco: SMOTE for Regression, *Progress in Artificial Intelligence*, Vol. 8154, pp. 378–389 (online), DOI: 10.1007/978-3-642-40669-0_33 (2013).
- [14] Paula Branco, Luís Torgo and Rita P. Ribeiro: SMOGN: A Pre-processing Approach for Imbalanced Regression, *Proceedings of 1st International Workshop on Learn-*

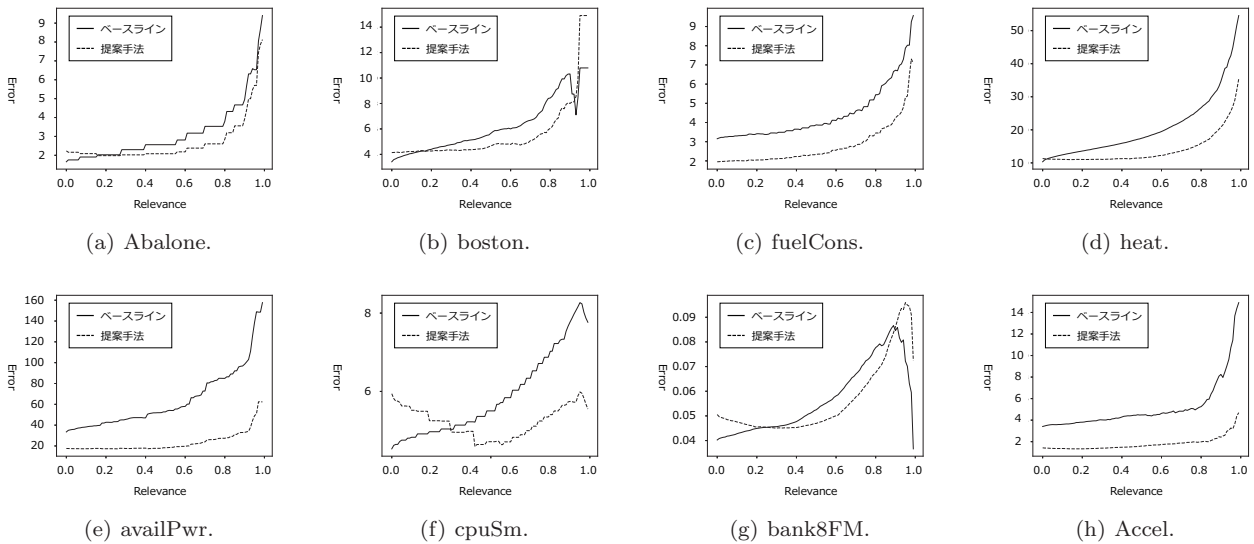


図 4 relevance function の値がしきい値以上の正解ラベルを持つデータに対する推定値の MAE. 各グラフは縦軸が MAE, 横軸がしきい値となる relevance function の値を示す. 提案手法を破線, ベースライン手法を実線で表す.

ing with Imbalanced Domains: Theory and Applications (2017).

- [15] Paula Branco, Luís Torgo and Rita P. Ribeiro: Pre-processing approaches for imbalanced distributions in regression, *Neurocomputing*, Vol. 343, pp. 76–99 (online), DOI: 10.1016/j.neucom.2018.11.100 (2019).
- [16] Luís Torgo and Rita Ribeiro: Utility-Based Regression, *Knowledge Discovery in Databases: PKDD 2007*, pp. 597–604 (2007).
- [17] Murray Rosenblatt: Remarks on Some Nonparametric Estimates of a Density Function, *The Annals of Mathematical Statistics*, Vol. 27, No. 3, pp. 832 – 837 (online), DOI: 10.1214/aoms/1177728190 (1956).
- [18] Emanuel Parzen: On Estimation of a Probability Density Function and Mode, *The Annals of Mathematical Statistics*, Vol. 33, No. 3, pp. 1065 – 1076 (online), DOI: 10.1214/aoms/1177704472 (1962).
- [19] David W. Scott: *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, Inc. (1992).
- [20] Luís Torgo and Ribeiro, R.: Predicting Rare Extreme Values, *Advances in Knowledge Discovery and Data Mining*, Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 816–820 (2006).