

不均衡データに対する多段階学習を用いた アンサンブルモデルによる2クラス分類アルゴリズムの提案

藤原 和樹^{1,a)} 繁野 麻衣子¹ 住田 潮¹

受付日 2019年1月31日, 再受付日 2019年3月21日,
採録日 2019年4月18日

概要: 近年, 機械学習を用いた2クラス分類アルゴリズムは, 種々のアプリケーション実現のための中核的な技術要素として応用されている. 一方で, 応用される多くの事例では対象データが少数の陽性と多数の陰性から構成される不均衡データであり, この不均衡性によって陽性の分類精度が低くなってしまうことが問題になっている. 本稿では, 不均衡データに対する陽性の分類精度の向上を目的とした2クラス分類アルゴリズムについて研究する. 一般的に, 不均衡データに対する陽性の分類精度を向上させようとした場合, 偽陰性の減少と引き換えに偽陽性が増加する可能性がある. そこで本研究では, 不均衡データに対して偽陰性・偽陽性をともに減少させる多段階学習を用いたアンサンブルモデルによる2クラス分類アルゴリズムを提案する. 提案手法では, 分類が難しいと判断されたデータに対して学習を繰り返し, 期ごとに複数のモデルを作成する. 多段階的に作成された複数のモデルを用いて, 偽陰性を減少させるための最適線形結合モデルと偽陽性を減少させるためのカスケード結合モデルをそれぞれ構築し, この2つのモデルを統合して最終的な分類を行う. 実験を通じて, 提案手法を用いることによって, 既存手法よりも陽性の分類精度が向上すること, 偽陰性と偽陽性がともに減少することを示した.

キーワード: 機械学習, 2クラス分類, 不均衡データ

A New Ensemble Model for Imbalanced Two-class Classification by Learning Multistage

KAZUKI FUJIWARA^{1,a)} MAIKO SHIGENO¹ USHIO SUMITA¹

Received: January 31, 2019, Revised: March 21, 2019,
Accepted: April 18, 2019

Abstract: In recent years, algorithms for classification with two possible outcomes have played important roles in machine learning. In various applications in the real world, the analyzing datasets are hard to deal with because the sizes of classes are imbalanced, i.e., the data contain a few positive outcomes while contain many negative outcomes. This imbalance causes the low accuracy of classification for positive. Many algorithms for two-class classification have been developed to improve the accuracy of classification even if the dataset is imbalanced. However, such algorithms tend to yield increasing false positives in return for reducing false negatives. We propose an algorithm tries to reduce both false negative and false positives in multistage learning. By using several learning machines, our algorithm constructs several models. To reduce false negatives, we take an optimal linear combination of those models. On the other hand, to reduce false positives, we take a cascade classifier. Integrating these two models, the solution for classification is given. Computational experiments are performed to verify the proposed method and to ascertain that the classification accuracy of positive is better than the existing method. The proposed method succeeded to decrease both the false negative and false positives.

Keywords: machine learning, two-class classification, imbalanced data

¹ 筑波大学
University of Tsukuba, Tsukuba, Ibaraki 305–8573, Japan
^{a)} s1720600@yahoo.co.jp

1. はじめに

第3次人工知能 (Artificial Intelligence: AI) ブームの到来により, AI に関する話題はメディアで取り上げられない日がないほど注目を浴びている. IoT (Internet of Things) 化による処理可能データの飛躍的増加や, 計算機の処理能力の向上, そして機械学習技術の進化によってこのブームが牽引されているといわれている [1]. なかでも現在, 機械学習を用いた2クラス分類はこのブームに相まって, 多岐にわたる分野で応用され始めている. たとえば, 医用画像診断やECサイトユーザのコンバージョン予測, クレジットカードの不正利用検出等があげられる. 応用事例では対象データが少数の陽性と多数の陰性から構成される不均衡なデータであることが多く, この不均衡性によって陽性の分類精度が低くなってしまふことが問題になっている. 本研究では, このような陽性が陰性に比べて極端に少ないデータを不均衡データと定義し, 不均衡データに対する2クラス分類に焦点を当てる.

Chawla [2] は, 不均衡データに対する分類精度が低下する問題によって, 実務へ悪影響を与える恐れがあると述べている. その1例として, CT画像に対してそれぞれ癌 (陽性)・健康 (陰性) のラベルが付与された Mammography dataset [3] を用いた医用画像診断を考える. このデータは260の陽性と10,923の陰性で構成された不均衡データになっている. 不均衡データであることを考慮せずに学習したモデルで分類した場合, ほとんどの患者を陰性と予測し, 大半の陽性を誤分類する結果になりやすい. 適合率が0~10%と低くなり, 特異度が90~100%と高くなる傾向にあった [4]. 適合率が10%のとき, 234人の患者の癌を見落とす誤診をしていることになり, 場合によっては病院の訴訟問題・信用問題にまで発展しかねない. このような不均衡データに対する分類精度の低下による弊害は, 医用画像診断だけではなく各分野共通の問題になっている [5]. したがって, 機械学習を用いた2クラス分類の応用が一層進むことが予想される今, 不均衡データに対する陽性の分類精度の向上は重要な解決すべき問題であるといえることができる.

そこで, 本稿では, 不均衡データに対する陽性の分類精度の向上を目的とした2クラス分類アルゴリズムについて研究する. 一般的に, 不均衡データに対する陽性の分類精度を向上させようとした場合, 偽陰性の減少と引き換えに偽陽性が増加してしまう問題が生じる. この問題を解決するための手法として, 不均衡データに対して偽陰性・偽陽性をともに減少させる多段階学習を用いたアンサンブルモデルによる2クラス分類アルゴリズムを提案する.

本研究の主要な貢献は以下のとおりである.

- 不均衡データに対する陽性の分類精度を向上させるために, 偽陰性・偽陽性をともに減少させる多段階学習を用いたアンサンブルモデルによる2クラス分類アル

ゴリズムを初めて提案した. 提案手法では, 分類が難しいと判断されたデータに対して学習を繰り返し, 期ごとに複数のモデルを作成する. 多段階的に作成された複数のモデルを用いて, 偽陰性を減少させるための最適線形結合モデルと偽陽性を減少させるためのカスケード結合モデルをそれぞれ構築し, この2つのモデルを統合して最終的な分類を行うことで陽性の分類精度向上を可能にしている.

- 実験を通じて, 最適線形結合モデル, 多段階学習, カスケード結合モデルそれぞれの有効性を検証し, これらを組み合わせた提案手法を用いることによって, 既存手法よりも陽性の分類精度が向上すること, 偽陰性と偽陽性がともに減少することを示した.

本研究の残りの構成は次のとおりである. 2章では, 不均衡データに対する2クラス分類アルゴリズムの研究動向についてまとめる. 3章では, 提案手法について説明を行い, 既存手法との関係について明らかにする. 4章では, 提案手法の有効性・実用性を確認するための検証実験の詳細および実験結果を説明する. 5章では, 本研究で取り組んだ不均衡データに対する陽性の分類精度の低下という問題への貢献および将来課題について述べる. 最後に6章で本稿をまとめる.

2. 関連研究

2.1 不均衡データに対する2クラス分類アルゴリズム

本章では, 不均衡データに対する2クラス分類アルゴリズムに関する主流なアプローチ [5] について紹介する. リサンプリング学習, アンサンブル学習, 両者を組み合わせたハイブリッドモデルの3つについて順に説明していく.

リサンプリング学習は, 学習データの陽性と陰性の比率が1:1になるようにリサンプリングしたデータを学習に用いる方法である [6]. このアプローチの大きな利点は, データの前処理段階に行うため, 様々な分類アルゴリズムと組み合わせることが可能な点にある [7]. リサンプリング学習のうち, 代表的なアンダサンプリングとオーバサンプリングについて説明する. アンダサンプリングは, 多数の陰性を少数の陽性と同数程度になるようにサンプリングを行う手法である [8]. 単純な分類精度の向上だけではなく, 学習データの不均衡性を取り除くことに加え, 陰性データ削減による計算コストの減少等のメリットがある. 一方で, 母集団の特徴を表すのに本来重要であったデータを取りこぼしてしまうデメリットがある. オーバサンプリングは, 少数の陽性を多数の陰性と同数程度になるようにサンプリングを行う手法である [8]. 単純な分類精度の向上に加え, 学習データの不均衡性を取り除くメリットがある一方で, 陽性データ増加による計算コストの増加や過学習のリスクがあるといったデメリットがある.

アンサンブル学習 [9] は, 複数の弱学習器を統合するこ

とによって分類精度を向上させる方法である。不均衡データに対して特に有効であるとされているバギングとブースティングについて説明する。バギングはブートストラップ法によって選ばれたデータ集合に対して学習を行い、構築された複数の弱学習器を統合する手法である。弱学習器間で異なるデータ集合を用いているため、予測結果のバリエーションが低下しやすいことや、学習を並行して行えるといった特徴がある。代表的な学習アルゴリズムとして、弱学習器を決定木としたランダムフォレストがあげられる。ブースティングは、学習データに対して逐次的に学習を行い、構築された複数の弱学習器を統合する手法である。1度誤分類したデータを正解できるように弱学習器を統合することによって、予測結果のバイアスが低くなりやすいといった特徴がある。代表的な学習アルゴリズムとして、誤分類したデータに対する重みが大きくなるように更新する Ada Boost, 損失関数が最小になるように弱学習器を構築する XGBoost や LightGBM [10] 等があげられる。

ハイブリッドモデル [11] は、データの预处理としてリサンプリングを行い、生成されたデータに対してアンサンブル学習を行う手法である。Wallace ら [12] は、確率論を用いてアンダサンプリングとバギングを組み合わせたハイブリッドモデルが最も有効な手法であると主張している。また、疑似データと実データを用いて他の手法との比較実験を行っており、主張をさらに強める結果となっていた。Salunkhe ら [13] は、アンダサンプリングとバギングをベースとしたモデルを提案し、複数用意された不均衡なデータセットに対して高い分類精度を実現した。このほかにもアンダサンプリングとバギングを組み合わせたハイブリッドモデルの有効性を示した研究は数多く報告されている [14]。

不均衡データに対して一般的な2クラス分類アルゴリズムを適用した場合、偽陽性は過少に、偽陰性は過多になる傾向がある [15]。紹介した手法は、過多であった偽陰性を減少させることで分類精度を向上させていることが多く見受けられる。しかし、偽陰性と偽陽性はトレードオフの関係にあるため、既存手法は偽陰性の減少の代償として本来過少であった偽陽性を増加させているともいえる。したがって、偽陰性・偽陽性の両者を同時に減少させることが、不均衡データに対する2クラス分類アルゴリズム開発の要諦になる。

3. 提案手法

提案手法は、分類が難しいと判断されたデータに対して学習を繰り返し、期ごとに複数のモデルを作成する。多段階的に作成された複数のモデルを用いて、偽陰性を減少させるための最適線形結合モデルと偽陽性を減少させるためのカスケード結合モデルをそれぞれ構築し、この2つのモデルを統合して最終的な分類を行うアンサンブルモデルになっている。本章では、はじめに、提案手法で行われる多

段階学習を用いたアンサンブルモデルについて説明し、次に、既存手法との関係と提案手法の新規性について述べる。

3.1 多段階学習を用いたアンサンブルモデル

多段階学習を用いたアンサンブルモデルでは、クラス1 (陽性) とクラス0 (陰性) に分類されている既知のデータ x の集合 D を考える。 D を学習データ集合 D_L と検証データ集合 D_V に分割する。それぞれのデータ集合に含まれるクラス i ($= 1, 0$) のデータ集合を $D_{L:i}$, $D_{V:i}$ と記す。 n ($= 1, 2, \dots$) 期でアンダサンプリングによって選ばれたデータ集合の族を M^n とし、多段階学習の n 期目で用いる学習データを D_L^n とする。ただし、1期目の学習データ D_L^1 は D_L を使用する。

n 期目では、はじめに、 D_L^n から選ばれた各データ集合 $D_m \in M^n$ に対して学習させた識別モデル $ALG_{n,D_m}(x)$ を作成する。次に、識別モデル $ALG_{n,D_m}(x)$ の重みを $\alpha_{n,D_m} \in [0, 1]$ とする最適線形結合モデル

$$LC_n(x) = \sum_{l=1}^n \sum_{D_m \in M^l} \alpha_{l,D_m} ALG_{l,D_m}(x) \quad (1)$$

を構築する。 α_{n,D_m} は、重みの総和が1になる条件の下で、 D_V に対する $LC_n(x)$ の分類精度 AUC-PR [16] を最大にするような重みとする。続いて、 D_V に対して再現率が1以上で適合率を最大にするように分類する識別関数

$$c_n(x) = \begin{cases} 1 & LC_n(x) \geq \underset{0 \leq \theta \leq 1}{\operatorname{argmax}} (Pre(x|\theta) : Rec(x|\theta) = 1) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

を弱学習機とするカスケード結合モデル

$$CC_n = \prod_{l=1}^n c_l(x), \quad (3)$$

を構築する。 $Pre(x|\theta)$ と $Rec(x|\theta)$ はそれぞれ D_V を閾値 $\theta \in [0, 1]$ で分類したときの適合率、再現率を表す。最適線形結合モデルとカスケード結合モデルを統合した n 期のアンサンブルモデルを $LC_n(x) \times CC_n(x)$ とし、各期のアンサンブルモデルの中で D_V に対する AUC-PR を最大にする l^* 期のモデル

$$F_{l^*}(x) = LC_{l^*}(x) \times CC_{l^*}(x) \quad (4)$$

を最終的な識別モデルとする。そして、 $n+1$ 期の学習データ D_L^{n+1} を $\{x \in D_L^n : c_n(x) = 1\}$ と更新して、 $n+1$ 期を同様に繰り返す。 k 回連続で l^* が変わらないとき、アルゴリズムを終了する。

3.2 既存手法との関係

提案手法は、2章で述べた不均衡データに対する2クラス分類アルゴリズムに関する主流なアプローチのうち、ハ

イブリッドモデルに該当する。ハイブリッドモデルの中で特に有効であるとされていたアンダサンプリングとバギングを組み合わせたモデルは、一般的に

$$EM(x) = \frac{\sum_{D_m \in M^1} ALG_{1,D_m}(x)}{|M^1|} \quad (5)$$

と表すことができる。提案手法には既存手法にはない特徴として、ハイブリッドモデルを多段階に学習を行うこと、最適線形結合モデル $LC_{l^*}^*(x)$ によって偽陰性の減少が可能になること、カスケード結合モデル $CC_{l^*}^*(x)$ によって偽陽性の減少が可能になることの3つがあげられる。既存手法 $EM(x)$ のようなハイブリッドモデルはこれまで数多く提案されてきた [14] 一方で、多段階に学習させる手法は筆者の調査時点で報告されていない。一般的なブースティングと多段階学習は、1度誤分類したデータを正解できるように学習させているという点で同じであるが、入力データの扱い方と最適線形結合する際の目的関数に違いがある。前者は入力データを学習データとし、学習データに対する予測残差の最小化を行っているのに対し、後者では入力データを学習データと検証データの2つに分割し、検証データに対する予測精度の最大化を行っている。したがって、多段階学習はブースティングよりも対象データの不均衡性を考慮しやすいかつ未知のデータに対する予測精度を向上させやすい学習方法といえる。

また、 $EM(x)$ は弱学習機を単純平均するのに対し、最適線形結合モデル $LC_{l^*}^*(x)$ は AUC-PR を最大にするように線形結合を行うため、比較的高い精度が期待できる。さらに、偽陰性の減少の代償として本来過少であった偽陽性を増加させているという多くの既存手法に関わる問題に対して、カスケード結合モデル $CC_{l^*}^*(x)$ を用いることで明らかな陰性に対して陽性と予測することを防ぐことが可能になり、偽陽性の減少が期待できる。

4. 実験

4.1 実験目的

提案手法の有効性・実用性に関する以下5点について検証および評価を行うことが本実験の目的である。(1) 最適線形結合モデルを用いることで既存手法よりも偽陰性が減少すること、(2) 多段階学習を用いたアンサンブルモデルの陽性の分類精度が通常の学習をさせたハイブリッドモデルよりも優れていること、(3) カスケード結合モデルを用いることで最適線形結合モデル単体で分類するときよりも偽陽性が少なくなること、(4) 既存手法と比較して提案手法の方が陽性の分類精度が優れていること、(5) 既存手法と比較して提案手法の計算時間が実用的な値であることを確認すること。

4.2 データセット

本実験では、オープンデータセットと実データセットの

2つを用いる。はじめに、オープンデータセットであるクレジットカードの利用履歴データについて、次に、某BtoC企業から提供されたECサイトのアクセスログデータについてそれぞれ説明していく。

クレジットカード利用履歴データ [17] を用いて、ある利用履歴データが不正取引（陽性）であるか正常な取引（陰性）であるかの分類を行う。陽性が492件、陰性が284,315件の計284,807件のデータで構築されており、全データに対する陽性割合は0.172%と極度な不均衡データになっている。特徴量は、主成分分析で変換済みの28変数と記録時間、そして取引金額の計30変数が与えられている。

ECサイトのアクセスログデータを用いて、あるセッションが2ページ目以降にECサイト会員に新規登録する（陽性）か新規登録しない（陰性）かの分類を行う。陽性が9,695件、陰性が3,050,818件で構築されており、全データに対する陽性割合は0.317%と極度な不均衡データになっている。特徴量は、ランディングページ情報のみを用いる。

4.3 使用する学習器と各パラメータ設定

本実験では、ロジスティック回帰、ニューラルネットワーク、LightGBM [10] の3つの学習器（以降LR, NN, LGB）を弱学習器として使用する。それぞれのハイパーパラメータは、Hyperopt [18] を用いて決定するものとする。提案手法のパラメータであるアンダサンプリングによって生成するデータ集合数 $|M^n|$ を50、アルゴリズム終了条件である k を3とする。

4.4 評価方法

不均衡データに対する分類精度の評価指標は、両クラスの精度のバランスを考慮する必要があるといわれている [19]。以上を踏まえ、検証実験に用いる評価指標をF値、AUC-PRとする。また、本研究では、偽陰性の減少と引き換えに偽陽性が増加する問題について議論するため、偽陰性と偽陽性にも着目する。

4.5 検証実験

本実験では、上記の目的を達成するため、次の方法により検証実験を行う。1) データセットを学習データ、検証データ、評価データをそれぞれ $|D_L| : |D_V| : |D_n| = 6 : 3 : 1$ かつ陽性割合が一定になるように層化抽出法を用いて3つに分割し、 D_L と D_V を用いてモデルを作成する。したがって分割された各データセットの特徴量の分布は一様ではない。2) D_T に対するモデルの分類精度を評価する。1), 2) の試行を50回繰り返した結果に対して有意水準1%でt検定を行い、2者間の差が統計的に有意か検証する。

4.6 実験1：最適線形結合モデルの有効性の検証

実験1では、平均モデルである既存手法 $EM(x)$ と提案

表 1 実験 1 結果：最適線形結合モデルの有効性の検証

Table 1 Result of experiment 1: effectiveness of optimal linear combination model.

学習器	評価指標	予測モデル	クレジットカードの利用履歴データ			ECサイトのアクセスログデータ		
			平均(SD)	t値	有意確率p	平均(SD)	t値	有意確率p
LR	AUC-PR	$EM(x)$	0.320 (0.024)	22.588	$p < .01$	0.338 (0.025)	4.140	$p < .01$
		$LC_1(x)$	0.45 (0.027)			0.36 (0.025)		
	F値	$EM(x)$	0.537 (0.033)	2.298	$p < .01$	0.527 (0.034)	3.633	$p < .01$
		$LC_1(x)$	0.553 (0.032)			0.548 (0.032)		
	偽陰性	$EM(x)$	20.28 (3.308)	2.486	$p < .01$	583.76 (15.905)	4.707	$p < .01$
		$LC_1(x)$	18.6 (3.58)			567.22 (18.665)		
偽陽性	$EM(x)$	69.74 (5.731)	12.241	$p < .01$	1178.34 (17.994)	10.393	$p < .01$	
	$LC_1(x)$	57.32 (4.391)			1140.26 (19.051)			
NN	AUC-PR	$EM(x)$	0.317 (0.026)	26.196	$p < .01$	0.359 (0.028)	3.158	$p < .01$
		$LC_1(x)$	0.465 (0.029)			0.378 (0.026)		
	F値	$EM(x)$	0.538 (0.032)	4.772	$p < .01$	0.544 (0.033)	3.963	$p < .01$
		$LC_1(x)$	0.569 (0.029)			0.569 (0.031)		
	偽陰性	$EM(x)$	19.02 (3.782)	2.948	$p < .01$	566.8 (17.624)	6.448	$p < .01$
		$LC_1(x)$	16.7 (3.43)			544.42 (18.631)		
偽陽性	$EM(x)$	68.92 (4.584)	15.518	$p < .01$	1163.96 (18.014)	11.841	$p < .01$	
	$LC_1(x)$	53.98 (5.081)			1123 (18.032)			
LGB	AUC-PR	$EM(x)$	0.325 (0.028)	24.183	$p < .01$	0.376 (0.024)	5.112	$p < .01$
		$LC_1(x)$	0.499 (0.035)			0.402 (0.028)		
	F値	$EM(x)$	0.548 (0.034)	8.979	$p < .01$	0.563 (0.031)	4.373	$p < .01$
		$LC_1(x)$	0.611 (0.038)			0.587 (0.03)		
	偽陰性	$EM(x)$	16.04 (3.27)	3.917	$p < .01$	556.94 (17.29)	9.410	$p < .01$
		$LC_1(x)$	13.8 (3.314)			522.06 (18.433)		
偽陽性	$EM(x)$	66.74 (4.261)	15.950	$p < .01$	1138.94 (17.386)	11.118	$p < .01$	
	$LC_1(x)$	50.94 (5.036)			1097.94 (18.64)			

手法の1期目の最適線形結合モデル $LC_1(x)$ の比較を行い、最適線形結合モデルの有効性を検証する。実験1の結果を表1に示す。表1は、両データセットに対して各学習器に学習させた既存手法 $EM(x)$ と提案手法の1期目の最適線形結合モデル $LC_1(x)$ の D_T に対する評価指標ごとの平均、標準偏差、検定結果を表している。 $LC_1(x)$ で予測した場合、両データともいずれの評価指標においても $EM(x)$ より分類精度が高くなる結果となった。 $EM(x)$ と $LC_1(x)$ の間に評価指標ごとの平均値の差が統計的に有意か確かめるために、有意水準1%で両側検定の t 検定を行った結果、二者間の評価指標ごとの差は有意であることが分かった。以上の結果から、最適線形結合モデルを用いることで既存手法よりも偽陰性と偽陽性をともに減少させることが確認することができた。

4.7 実験2：多段階学習の有効性の検証

実験2では、提案手法の1期目の識別モデル $F_1(x)$ と l^* 期目の識別モデル $F_{l^*}(x)$ の比較を行い、多段階学習の有効性を検証する。実験2の結果を表2に示す。表2は、両データセットに対して各学習器に学習させた提案手法の1期目の識別モデル $F_1(x)$ と l^* 期目の識別モデル $F_{l^*}(x)$ の D_T に対する評価指標ごとの平均、標準偏差、検定結果を表している。 $F_{l^*}(x)$ で予測した場合、両データともいずれの評価指標においても $F_1(x)$ より分類精度が高くなる結果となった。 $F_1(x)$ と $F_{l^*}(x)$ の間に評価指標ごとの平均値の差が統計的に有意か確かめるために、有意水準1%で両側検定の t 検定を行った結果、二者間の評価指標ごとの差は有意であることが分かった。以上の結果から、多段階学習

表 2 実験 2 結果：多段階学習の有効性の検証

Table 2 Result of experiment 2: effectiveness of multistage learning.

学習器	評価指標	比較対象	クレジットカードの利用履歴データ			ECサイトのアクセスログデータ		
			平均(SD)	t値	有意確率p	平均(SD)	t値	有意確率p
LR	AUC-PR	$F_1(x)$	0.454 (0.029)	40.805	$p < .01$	0.361 (0.027)	25.659	$p < .01$
		$F_{l^*}(x)$	0.741 (0.039)			0.488 (0.029)		
	F値	$F_1(x)$	0.578 (0.035)	34.793	$p < .01$	0.568 (0.03)	12.918	$p < .01$
		$F_{l^*}(x)$	0.844 (0.033)			0.635 (0.027)		
	偽陰性	$F_1(x)$	18.56 (3.621)	4.232	$p < .01$	565.4 (17.795)	52.138	$p < .01$
		$F_{l^*}(x)$	15.46 (3.765)			393.88 (18.073)		
偽陽性	$F_1(x)$	48.04 (4.145)	40.171	$p < .01$	1139.48 (16.522)	13.440	$p < .01$	
	$F_{l^*}(x)$	15.4 (3.446)			1094.56 (17.544)			
NN	AUC-PR	$F_1(x)$	0.473 (0.028)	46.008	$p < .01$	0.38 (0.027)	22.821	$p < .01$
		$F_{l^*}(x)$	0.751 (0.036)			0.516 (0.034)		
	F値	$F_1(x)$	0.613 (0.035)	35.650	$p < .01$	0.591 (0.033)	10.538	$p < .01$
		$F_{l^*}(x)$	0.866 (0.03)			0.653 (0.025)		
	偽陰性	$F_1(x)$	15.92 (3.59)	3.844	$p < .01$	538.12 (16.442)	47.77	$p < .01$
		$F_{l^*}(x)$	13 (3.136)			375.28 (17.585)		
偽陽性	$F_1(x)$	46.46 (4.161)	46.051	$p < .01$	1119.82 (16.603)	14.565	$p < .01$	
	$F_{l^*}(x)$	12.96 (3.05)			1071.7 (17.277)			
LGB	AUC-PR	$F_1(x)$	0.513 (0.031)	33.366	$p < .01$	0.407 (0.028)	17.493	$p < .01$
		$F_{l^*}(x)$	0.741 (0.042)			0.522 (0.032)		
	F値	$F_1(x)$	0.639 (0.032)	28.293	$p < .01$	0.616 (0.033)	10.227	$p < .01$
		$F_{l^*}(x)$	0.857 (0.037)			0.674 (0.027)		
	偽陰性	$F_1(x)$	13.9 (3.012)	3.905	$p < .01$	526.26 (17.784)	50.484	$p < .01$
		$F_{l^*}(x)$	11.26 (3.102)			350.14 (17.764)		
偽陽性	$F_1(x)$	42.92 (4.485)	41.448	$p < .01$	1095.08 (17.47)	12.509	$p < .01$	
	$F_{l^*}(x)$	10.84 (3.139)			1048.04 (17.099)			

を用いたアンサンブルモデルが通常の学習をさせたハイブリッドモデルよりも陽性の分類精度が優れていることを確認することができた。

4.8 実験3：カスケード結合モデルの有効性の検証

カスケード結合モデルを用いることによって、偽陽性が減少することを確認する。実験3では、識別モデル $LC_{l^*}(x)$ と識別モデル $F_{l^*}(x)$ ($= LC_{l^*}(x) \times CC_{l^*}(x)$) の比較を行い、カスケード結合モデルの有効性の検証する。実験3の結果を表3に示す。表3は、両データセットに対して各学習器に学習させた識別モデル $LC_{l^*}(x)$ と識別モデル $F_{l^*}(x)$ の D_T に対する評価指標ごとの平均、標準偏差、検定結果を表している。 $F_{l^*}(x)$ で予測した場合、両データともいずれの評価指標においても $LC_{l^*}(x)$ より分類精度が高くなる結果となった。 $LC_{l^*}(x)$ と $F_{l^*}(x)$ の間に評価指標ごとの平均値の差が統計的に有意か確かめるために、有意水準1%で両側検定の t 検定を行った結果、二者間の評価指標ごとの差は有意であることが分かった。以上の結果から、カスケード結合モデルを用いることで最適線形結合モデル単体で分類するときよりも偽陽性が少なくなることを確認することができた。

4.9 実験4：提案手法の有効性の検証

実験4では、既存手法モデル $EM(x)$ と提案手法モデル $F_{l^*}(x)$ の比較を行い、その有効性を検証する。実験4の結果を表4に示す。表4は、両データセットに対して各学習器に学習させた既存手法モデル $EM(x)$ と提案手法モデル $F_{l^*}(x)$ の D_T に対する評価指標ごとの平均、標準偏差、

表 3 実験 3 結果：カスケード結合モデルの有効性の検証

Table 3 Result of experiment 3: effectiveness of cascade classifier model.

学習器	評価指標	比較対象	クレジットカードの利用履歴データ			ECサイトのアクセスログデータ		
			平均(SD)	t値	有意確率p	平均(SD)	t値	有意確率p
LR	AUC-PR	$LC_{\tau}(x)$	0.711 (0.037)	3.667	$p < .01$	0.459 (0.027)	4.246	$p < .01$
		$F_{\tau}(x)$	0.74 (0.039)			0.483 (0.03)		
	F値	$LC_{\tau}(x)$	0.764 (0.039)	12.789	$p < .01$	0.61 (0.03)	4.257	$p < .01$
		$F_{\tau}(x)$	0.859 (0.03)			0.635 (0.03)		
	偽陰性	$LC_{\tau}(x)$	18.88 (3.662)	4.144	$p < .01$	542.86 (19.976)	37.059	$p < .01$
		$F_{\tau}(x)$	15.92 (3.816)			399.68 (20.257)		
	偽陽性	$LC_{\tau}(x)$	34.86 (3.969)	22.271	$p < .01$	1109.04 (18.282)	5.701	$p < .01$
		$F_{\tau}(x)$	15.84 (3.639)			1088.72 (18.045)		
NN	AUC-PR	$LC_{\tau}(x)$	0.711 (0.037)	4.043	$p < .01$	0.479 (0.028)	5.383	$p < .01$
		$F_{\tau}(x)$	0.738 (0.035)			0.509 (0.031)		
	F値	$LC_{\tau}(x)$	0.787 (0.036)	11.292	$p < .01$	0.642 (0.036)	2.207	$p < .01$
		$F_{\tau}(x)$	0.859 (0.03)			0.656 (0.027)		
	偽陰性	$LC_{\tau}(x)$	16.6 (3.064)	2.894	$p < .01$	521.26 (18.736)	43.388	$p < .01$
		$F_{\tau}(x)$	14.8 (3.283)			376.94 (15.953)		
	偽陽性	$LC_{\tau}(x)$	33.36 (4.332)	23.427	$p < .01$	1097.66 (17.648)	7.361	$p < .01$
		$F_{\tau}(x)$	14.74 (3.122)			1073.76 (17.318)		
LGB	AUC-PR	$LC_{\tau}(x)$	0.72 (0.038)	5.216	$p < .01$	0.492 (0.029)	5.553	$p < .01$
		$F_{\tau}(x)$	0.75 (0.036)			0.531 (0.033)		
	F値	$LC_{\tau}(x)$	0.823 (0.041)	5.915	$p < .01$	0.653 (0.042)	2.693	$p < .01$
		$F_{\tau}(x)$	0.86 (0.03)			0.671 (0.027)		
	偽陰性	$LC_{\tau}(x)$	13.24 (2.847)	3.396	$p < .01$	499.12 (18.272)	40.204	$p < .01$
		$F_{\tau}(x)$	11.22 (3.066)			348.04 (16.919)		
	偽陽性	$LC_{\tau}(x)$	29.58 (3.959)	23.612	$p < .01$	1076.16 (17.702)	7.728	$p < .01$
		$F_{\tau}(x)$	11.26 (2.863)			1050.72 (18.306)		

表 4 実験 4 結果：提案手法の有効性の検証

Table 4 Result of experiment 4: effectiveness of the proposed method.

学習器	評価指標	比較対象	クレジットカードの利用履歴データ			ECサイトのアクセスログデータ		
			平均(SD)	t値	有意確率p	平均(SD)	t値	有意確率p
LR	AUC-PR	$EM(x)$	0.322 (0.026)	67.488	$p < .01$	0.343 (0.028)	25.373	$p < .01$
		$F_{\tau}(x)$	0.748 (0.038)			0.477 (0.025)		
	F値	$EM(x)$	0.535 (0.033)	45.619	$p < .01$	0.52 (0.033)	17.43	$p < .01$
		$F_{\tau}(x)$	0.847 (0.032)			0.638 (0.032)		
	偽陰性	$EM(x)$	20.28 (3.529)	6.683	$p < .01$	592.74 (15.043)	60.134	$p < .01$
		$F_{\tau}(x)$	15.94 (3.425)			398.22 (17.819)		
	偽陽性	$EM(x)$	68.74 (4.707)	64.757	$p < .01$	1176.86 (15.64)	24.057	$p < .01$
		$F_{\tau}(x)$	15.12 (3.192)			1090.08 (18.792)		
NN	AUC-PR	$EM(x)$	0.326 (0.026)	64.850	$p < .01$	0.359 (0.027)	24.743	$p < .01$
		$F_{\tau}(x)$	0.74 (0.038)			0.514 (0.036)		
	F値	$EM(x)$	0.54 (0.034)	46.749	$p < .01$	0.536 (0.032)	19.161	$p < .01$
		$F_{\tau}(x)$	0.849 (0.033)			0.653 (0.029)		
	偽陰性	$EM(x)$	18.48 (3.471)	5.432	$p < .01$	572.26 (16.455)	62.267	$p < .01$
		$F_{\tau}(x)$	14.52 (3.215)			379.14 (18.751)		
	偽陽性	$EM(x)$	69.18 (5.333)	57.850	$p < .01$	1161.94 (18.371)	28.463	$p < .01$
		$F_{\tau}(x)$	14.28 (3.338)			1066.82 (17.646)		
LGB	AUC-PR	$EM(x)$	0.326 (0.025)	61.926	$p < .01$	0.376 (0.024)	27.434	$p < .01$
		$F_{\tau}(x)$	0.742 (0.038)			0.53 (0.031)		
	F値	$EM(x)$	0.545 (0.03)	55.291	$p < .01$	0.572 (0.032)	16.269	$p < .01$
		$F_{\tau}(x)$	0.861 (0.03)			0.677 (0.026)		
	偽陰性	$EM(x)$	14.06 (3.467)	4.371	$p < .01$	548.62 (17.206)	57.032	$p < .01$
		$F_{\tau}(x)$	11.4 (3.194)			351.94 (19.326)		
	偽陽性	$EM(x)$	64.88 (4.868)	67.935	$p < .01$	1137.72 (15.468)	26.552	$p < .01$
		$F_{\tau}(x)$	10.6 (2.893)			1051.98 (17.159)		

検定結果を表している。 $F_{\tau^*}(x)$ で予測した場合、両データともいずれの評価指標においても $EM(x)$ より分類精度が高くなる結果となった。 $EM(x)$ と $F_{\tau^*}(x)$ の間に評価指標ごとの平均値の差が統計的に有意か確かめるために、有意水準 1% で両側検定の t 検定を行った結果、二者間の評価指標ごとの差は有意であることが分かった。以上の結果から、既存手法と比較して提案手法の方が陽性の分類精度が優れていることを確認することができた。

表 5 実験 5 結果：提案手法の実用性の評価

Table 5 Result of experiment 5: performance of the proposed method.

学習器	比較対象	クレジットカードの利用履歴データ				ECサイトのアクセスログデータ			
		計算時間(s)		l^*		計算時間(s)		l^*	
		平均(SD)	Max (min)	平均(SD)	Max (min)	平均(SD)	Max (min)	平均(SD)	Max (min)
LR	既存手法	12.075 (0.305)	13.485 (11.859)	/	/	103.971 (1.204)	107.024 (100.512)	/	/
	提案手法	222.445 (115.755)	388.296 (94.919)	9.100 (3.422)	21 (5)	1134.978 (218.295)	1636.239 (550.058)	10.250 (4.147)	17 (5)
NN	既存手法	32.554 (0.632)	34.129 (31.602)	/	/	324.884 (3.124)	330.876 (316.150)	/	/
	提案手法	635.034 (406.171)	1852.805 (332.828)	9.720 (4.102)	25 (5)	6429.428 (1190.210)	9134.546 (1990.210)	8.920 (3.849)	16(6)
LGB	既存手法	37.471 (0.239)	37.986 (37.077)	/	/	341.986 (1.173)	345.211 (339.080)	/	/
	提案手法	392.503 (211.356)	1269.639 (167.934)	8.480 (3.829)	18 (5)	2956.603 (458.252)	3569.639 (1617.934)	8.480 (3.829)	18 (5)

4.10 実験 5：提案手法の実用性の検証

実験 5 では、既存手法と提案手法の計算時間の比較実験を行い、提案手法の実用性の評価する。実験 5 の結果を表 5 に示す。

表 5 は、両データセットに対する各学習器ごとの既存手法と提案手法の計算時間と l^* の平均・標準偏差・最大値・最小値を表している。提案手法では既存手法に対して多段階学習を行っていることから、計算時間は必然的に多くなるため、検定結果は割愛する。提案手法の平均計算時間は既存手法と比較して 3 倍から 20 倍の間で増加しており、クレジットカードの利用履歴データでは最大で約 30 分 52 秒、EC サイトのアクセスログデータでは最大で約 2 時間 32 分 14 秒かかる結果となった。また、 l^* の平均は学習器およびデータセットによらず 10 前後になっているが、5 から 25 の間で推移しており、安定性が低いことが確認された。

5. 考察

不均衡データに対する陽性の分類精度の低下という問題に対する提案手法の有効性と実用性について既存手法と比較しながら考察をする。実験 1 から、アンダサンプリングとバギングを組み合わせたハイブリッドモデルでは、統合方法を単純平均ではなく最適線形結合にすることで精度が向上することが確認された。既存手法のような単純平均による統合はすべてのモデルを等質に扱っているため精度の低いモデルの影響を受けやすいのに対し、最適線形結合モデルは、その影響を受けにくい分類精度が高くなったと考えられる。

実験 2 では、多段階学習を用いることによって、通常の学習させたモデルよりも精度が高くなることが確認された。多段階学習は、分類が難しいと判断されたデータに対して学習を繰り返すため、母集団の特徴を表すのに本来重要であったデータを取りこぼしてしまう可能性が比較的低い。加えて、期を重ねるごとに統合するモデルの数が増えることも精度向上につながった要因として考えられた。

実験 3 の結果から、最適線形結合モデルはカスケード結合モデルと統合させることで精度が向上することが確認された。カスケード結合モデルが陽性と判断したデータに対

してのみ最適線形結合モデルで予測を行うため、偽陽性が生じる可能性が減り、結果として精度が向上したと考えられる。

実験4では、不均衡データに対する陽性の分類精度の低下という問題に対して、既存手法よりも提案手法が有効であることが確認された。既存手法が陥っていた偽陰性の減少と引き換えに偽陽性が増加してしまう問題に対しても、比較的偽陰性・偽陽性をともに減少させることに成功していた。異なるデータセットに対して実験を繰り返したことで、複数の学習器を用いたこと、統計的手法を用いて検証したことによって、提案手法の汎用性が高いことが示唆された。

実験5の結果から、提案手法は既存手法よりも計算時間が大幅に増加するもののモデルの更新頻度が一定以下の場合に限り実用が可能であることを示唆された。実験データセットと同程度のデータサイズの時、クレジットカード不正利用検出モデルは31分に1回以下、ECサイトのコンバージョン予測は2時間33分に1回以下の頻度で更新する場合、実用可能であることが確認された。

次に、各データセットに対する提案手法の誤分類が実務に対してどれだけの影響を与えるかを議論するために、それぞれの応用事例にモデルを適用した場合を考える。

クレジットカードの不正利用検出にモデルを適用した場合を考える。クレジットカード利用履歴データでは、不正利用検出を行っているため、予測結果を基に不正利用者の検出に用いられる。したがって、偽陰性は見逃してしまったクレジットカードの不正利用件数と等しく考えることができる。一般的に、不正利用により顧客の失った金額は盗難保険等によって賄われ、クレジットカード会社が負担することになっている。つまり、クレジットカード会社にとって偽陰性の増加は、金銭的費用増加と信用低下につながるという。また、偽陽性は不正利用したと濡れ衣を着せられた利用件数と等しく考えることができる。つまり、クレジットカード会社にとって偽陽性の増加は、信用低下につながるという。したがって、提案手法は、信用問題に発展する要因である偽陰性と偽陽性を完全に排除できていないため、実務に応用する場合、さらなる精度改善が求められることが考えられる。

ECサイトのアクセスログデータでは、セッションの新規会員登録予測を行っているため、予測結果を基にWeb広告配信におけるターゲットセグメントの作成に用いられることが予想される。したがって、偽陰性は見逃してしまった新規登録見込みセッション数と等しく考えることができる。つまり、某BtoC企業にとって偽陰性の増加は、会員登録したであろうセッションを見逃す機会損失につながるという。また、偽陽性はターゲットセグメントのなかで会員登録をしないセッション数と考えることができる。つまり、某BtoC企業にとって偽陽性の増加は、Web広告配

信の費用対効果の低下につながるという。したがって、提案手法は、Web広告の効率が低下する要因である偽陰性と偽陽性を完全に排除できていないため、実務に応用する場合、さらなる精度改善が求められることが考えられる。

以上から、頻繁にモデルを更新する必要がある事例に対しても応用できるように高速化を図ること、誤分類の重みが特に大きい事例に対しても応用できるように高精度化すること等が今後の課題としてあげられる。

6. おわりに

機械学習を用いた2クラス分類アルゴリズムは、種々のアプリケーション実現のために応用されており、重要な技術となっている。しかし、実務への応用にあたって、不均衡データに対する陽性の分類精度の低下は大きな問題になっている。既存手法を用いることで一定の解決は見られるものの、偽陰性の減少と引き換えに偽陽性を増加させてしまうという課題がある。本稿ではこの問題と残された課題に対して、多段階学習を用いたハイブリッドモデルによる2クラス分類アルゴリズムを提案した。提案手法では、分類が難しいと判断されたデータに対して学習を繰り返し、期ごとに複数のモデルを作成する。多段階的に作成された複数のモデルを用いて、偽陰性を減少させるための最適線形結合モデルと偽陽性を減少させるためのカスケード結合モデルを構築し、この2つのモデルを統合して最終的な分類を行う方法である。実験を通じて、弱学習機の統合方法を線形結合にすることの有効性、多段階学習の有効性、カスケード結合モデルの有効性、提案手法を用いることによって既存手法よりも陽性の分類精度が向上し、偽陰性と偽陽性をともに減少することを示した。今後は、より不均衡なデータセットでの検証やより実務環境に近い問題設定において実験を行うことで、機械学習を用いた2クラス分類アルゴリズムの実務への応用に貢献していきたい。

参考文献

- [1] He, H. and Garcia, E.A.: Learning from imbalanced data, *IEEE Trans. Knowledge & Data Engineering*, Vol.21, No.9, pp.1263-1284 (2009).
- [2] Chawla, N.V.: Data mining for imbalanced datasets: An overview, *Data Mining and Knowledge Discovery Handbook*, pp.853-867, Springer (2005).
- [3] Woods, K.S., Doss, C.C., Bowyer, K.W., Solka, J.L., Priebe, C.E. and Kegelmeyer, Jr., W.P.: Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography, *International Journal of Pattern Recognition and Artificial Intelligence*, Vol.7, No.6, pp.1417-1436 (1993).
- [4] Chen, K., Lu, B.-L. and Kwok, J.T.: Efficient classification of multi-label and imbalanced data using min-max modular classifiers, *IJCNN*, pp.1770-1775 (2006).
- [5] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. and Bing, G.: Learning from class-imbalanced data: Review of methods and applications,

- Expert Systems with Applications*, Vol.73, pp.220–239 (2017).
- [6] Chawla, N.V., Japkowicz, N. and Kotcz, A.: Special issue on learning from imbalanced data sets, *ACM Sigkdd Explorations Newsletter*, Vol.6, No.1, pp.1–6 (2004).
- [7] Lane, P.C.R., Clarke, D. and Hender, P.: On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data, *Decision Support Systems*, Vol.53, No.4, pp.712–718 (2012).
- [8] Kotsiantis, S., Kanellopoulos, D. and Pintelas, P.: Handling imbalanced datasets: A review, *GESTS International Trans. Computer Science and Engineering*, Vol.30, No.1, pp.25–36 (2006).
- [9] Dietterich, T.G.: Ensemble methods in machine learning, *International Workshop on Multiple Classifier Systems*, pp.1–15, Springer (2000).
- [10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y.: LightGBM: A highly efficient gradient boosting decision tree, *Advances in Neural Information Processing Systems*, pp.3146–3154 (2017).
- [11] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F.: A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches, *IEEE Trans. Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol.42, No.4, pp.463–484 (2012).
- [12] Wallace, B.C., Small, K., Brodley, C.E. and Trikalinos, T.A.: Class imbalance, redux, *IEEE 11th International Conference on Data Mining*, pp.754–763 (2011).
- [13] Salunkhe, U.R. and Mali, S.N.: Classifier ensemble design for imbalanced data classification: A hybrid approach, *Procedia Computer Science*, Vol.85, pp.725–732 (2016).
- [14] Yang, Y., Xiao, P., Cheng, Y., Liu, W. and Huang, Z.: Ensemble strategy for hard classifying samples in class-imbalanced data set, *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp.170–175, IEEE (2018).
- [15] Ali, A., Shamsuddin, S.M. and Ralescu, A.L.: Classification with class imbalance problem: A review, *Int. J. Adv. Soft. Comput. Appl.*, Vol.7, No.3, pp.176–204 (2015).
- [16] Davis, J. and Goadrich, M.: The relationship between precision-recall and ROC curves, *Proc. 23rd International Conference on Machine Learning*, pp.233–240, ACM (2006).
- [17] Credit Card Fraud Detection Anonymized credit card transactions labeled as fraudulent or genuine, available from (<https://www.kaggle.com/mlg-ulb/creditcardfraud>).
- [18] Bergstra, J., Yamins, D. and Cox, D.D.: Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms, *Proc. 12th Python in Science Conference*, pp.13–20, Citeseer (2013).
- [19] Krawczyk, B.: Learning from imbalanced data: Open challenges and future directions, *Progress in Artificial Intelligence*, Vol.5, No.4, pp.221–232 (2016).



藤原 和樹

筑波大学大学院システム情報工学研究科社会学専攻修士課程修了。機械学習、データマイニングの研究に従事。



繁野 麻衣子

東京工業大学大学院理工学研究科情報科学専攻博士後期課程退学。博士（理学）。組合せ最適化，オペレーションズ・リサーチの研究に従事。



住田 潮

ロチェスター大学経営大学院博士課程修了。博士（Ph.D.）。応用数学，ビッグデータアナリティクス，マーケティング・サイエンスの研究に従事。