Vol.2018-MPS-121 No.10
2018/12/18

不均衡データに対する多段階学習を用いた 2クラス分類アルゴリズムの提案とその検証

藤原　和樹[1]　繁野 麻衣子[1]　住田 潮[1]

概要：EC サイトユーザーのコンバージョン予測等，機械学習を用いた 2 クラス分類は様々な事例において応用されている．その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データであり，この不均衡性によって識別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，このようなデータを不均衡データと定義し，不均衡データに対する分類精度向上を目的とした多段階の 2 クラス分類アルゴリズムを提案する．提案アルゴリズムでは，分類が難しいと判断されたデータに対してアンダーサンプリングとバギングを組み合わせたモデルの学習を繰り返し，多段階的に構築された複数のモデルを統合して分類を行う．そして，実データに対し提案アルゴリズムと既存アルゴリズムの比較実験を行い，有効性を検証する．

キーワード：機械学習，2 クラス分類問題，不均衡データ

1. はじめに

近年，機械学習を用いた 2 クラス分類は様々な事例において応用されている．例えば，EC サイトユーザーのコンバージョン予測やスパムメール検出，医療診断等が挙げられる．機械学習が幅広い分野での問題解決に有効な手段として，今後一層浸透していくことが予想される．一方で，その多くの事例では対象データの陽性・陰性の不均衡性によって分類精度を低下させてしまうことが問題になっている [1]．本研究では，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分類精度向上を目的とした 2 クラス分類アルゴリズムを提案する．

2. 不均衡データ分類問題の関連研究

現在では，不均衡データに対する学習方法として，リサンプリング学習，アンサンブル学習，そして両者を組み合わせたハイブリッドモデルの 3 つが主流になっている [2]．リサンプリング学習は，学習データを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．多数の陰性を減らすアンダーサンプリング，小数の陽性を増やすオーバーサンプリング等が挙げられる．アンサンブル学習は，ブートストラップサンプリングを繰り返して生成した弱学習機を統合するバギングと逐次的に弱学習機を統合するブースティング等が挙げられる．Salunkhe and Mali[2] は，アンダーサンプリングとバギングのハイブリッドモデルを用いることによって，上記の学習方法よりも分類精度が高い結果を示した．この他にもハイブリッドモデルの有効性を示した研究が多く存在する一方で，多段階に学習・分類する拡張手法は，我々の調査時点で報告されていない．そこで本研究では，不均衡データに対する分類精度向上を目的とした多段階の 2 クラス分類アルゴリズムを提案する．

2 クラス分類問題に対する精度評価指標は，全体の正解率を表す Accuracy を用いることが一般的であった．しかし，偽陽性と偽陰性を区別していないという問題があり，この問題こそが不均衡データに対する分類精度を低下させる要因の 1 つになっている．したがって，不均衡データに対する分類精度を評価する際には，両クラスの正解率のバランスを考慮する必要がある．現在では，分類精度を評価する場合には F1 値や G-mean が，予測確率の精度を評価する場合には AUC-ROC や AUC-PR が用いられることが多くなっている [3]．

3. 提案アルゴリズム

クラス 1（陽性）とクラス 0（陰性）に分類されているデータ x の集合がある．このデータ集合を学習データ D_L と検証データ D_V に分割する．それぞれのデータに含まれ

[1] 筑波大学

るクラス $i(i = 1, 0)$ のデータ集合を $D_{L:i}$，$D_{V:i}$ と記す．使用する学習器集合を K，アンダーサンプリングによって選ばれたデータ集合の族を M とする．

提案アルゴリズムでは，期ごとに識別モデルと識別関数を作成する．前の期で分類が難しいと判断されたデータを用いて次の期の識別モデルを作成する．1 期目では，与えられた学習データ D_L と検証データ D_V を用いる．提案アルゴリズムの $n(n = 1, 2, ...)$ 期目で用いる学習データと検証データを各々 D_L^n, D_V^n とする．1 期目では，与えられた学習データ DL と検証データ DV を用いる．つまり，$D_L^1 = D_L, D_V^1 = D_V$ とする．n 期目では，はじめに，D_L^n から選ばれた各データ集合 $D_m \in M$ に対して，各学習器 $k \in K$ を用いて学習を行い，識別モデル $ALG_{m,k}(x)$ を作成し，この平均 $P_n(x) := \sum_{D_m \in M, k \in K} ALG_{m,k}(x)/|M||K|$ を n 期の識別モデルとする．この識別モデル $P_n(x)$ を用いて D_V^n を閾値 $z \in [0, 1]$ で分類したときの識別関数 $I(x|z)$ に対する Precision と Recall を $Pr(x|z)$，$Re(x|z)$ で表す．ここで，$z^\star = \underset{0 \le z \le 1}{\mathrm{argmax}}(Pr(x|z)|Re(x|z) = 1)$ を求め，n 期の識別関数を $C_n(x) = I(x|z^\star)$ とする．次に $n + 1$ 期の学習データと検証データ $D_L^{n_1}, D_V^{n+1}$ を $\{x \in D_L^n | C_n(x) = 1\}$，$\{x \in D_V^n | C_n(x) = 1\}$ と更新して，$n + 1$ 期を同様に繰り返す．l 期と $l + 1$ 期で検証データ D_V が変わらないとき ($D_V^n = D_V^{n+1}$)，アルゴリズムを終了し，最終的な識別モデルと識別関数を

$$P_l(x) \prod_{n=1}^{l} C_n(x), \tag{1}$$

$$P_l(x|z^\star) \prod_{n=1}^{l} C_n(x) \tag{2}$$

で与える．ただし，$z^\star$ は l 期で求めた閾値である．

4. 提案アルゴリズムの評価

4.1 評価方法

提案アルゴリズムの有効性を評価するために，既存アルゴリズムであるアンダーサンプリングとバギングのハイブリッドモデル [2] との比較実験を行う．評価指標は 2 章を踏まえ，分類精度の評価に F1 値と G-mean，予測確率の精度の評価に AUC-ROC と AUC-PR を用いる．

本研究では，オープンデータのクレジットカード利用履歴データを用いた不正利用検出，某企業から提供された EC サイトの 2018 年 5 月から 8 月のアクセスログデータを用いたコンバージョン予測の 2 つの実験を行った．全データに対する陽性割合は，それぞれ 0.002，0.003 と不均衡性の高いデータになっている．共通条件として，学習器集合 K はロジスティク回帰，ランダムフォレスト，勾配ブースティング木の 3 つ，アンダーサンプリングにより生成するデータ集合の数 $|M|$ を 100，D_V に対して $F1$ 値を最大にする $z_n^\star$ を採用する．

4.2 結果

不正利用検出の結果を表 1 に，コンバージョン予測の結果を表 2 に示す．提案アルゴリズムにより分類を行った場合，両データともいずれの評価指標においても比較アルゴリズムよりも精度が向上した．偽陽性数はそれぞれ比較アルゴリズムの 9.5%，64.2% になっており，偽陽性数の大幅な減少が精度の向上に繋がっているといえる．

表 1 不正利用検出の比較実験結果

	比較アルゴリズム			提案アルゴリズム		
混同行列		正解			正解	
		陽性	陰性		陽性	陰性
予測	陽性	42	63	予測 陽性	42	6
	陰性	8	28,369	予測 陰性	8	28,426
$F1$値	0.542			0.857		
$G-mean$	0.632			0.935		
$AUC-ROC$	0.909			0.912		
$AUC-PR$	0.337			0.734		

表 2 コンバージョン予測の比較実験結果

	比較アルゴリズム			提案アルゴリズム		
混同行列		正解			正解	
		陽性	陰性		陽性	陰性
予測	陽性	1,344	548	予測 陽性	1,305	352
	陰性	1,138	675,872	予測 陰性	1,177	677,245
$F1$値	0.614			0.631		
$G-mean$	0.842			0.887		
$AUC-ROC$	0.762			0.770		
$AUC-PR$	0.390			0.420		

5. おわりに

本稿では，不均衡データに対する分類精度向上を目的とした分類アルゴリズムの提案を行い，その有効性について実データを用いて検証をした．一定の有効性が示唆されたものの，提案アルゴリズムをより多くのデータセットに対して適用し，汎用性について考察することが今後の課題としてあげられる．

参考文献

[1] Fernández, A., del Río, S., Chawla, N. V. and Herrera, F.: An insight into imbalanced Big Data classification: outcomes and challenges, *Complex & Intelligent Systems* (2017).
[2] 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).
[3] 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).