1M-05

データの複雑性を考慮した 不均衡データに対するリサンプリング手法

笹田 大翔† 馬場 睦也‡ 木村 優介‡ 波多野 賢治†
†同志社大学 文化情報学部 ‡同志社大学大学院 文化情報学研究科

1 はじめに

クラスの構成要素に偏りのある不均衡なデータを機械学習に適用した場合，予測結果は悪影響を受けてしまう．こうした偏りを解消する方法として，各クラス内のサンプル数を調整するリサンプリングという手法が用いられる．しかしクラスの偏り以外にも学習に影響を与えるものとして，学習時に他クラスに分類されてしまう特徴量をもつ要素が存在するノイズや，クラス同士が重複して線形に分離できない状態を表すオーバーラップといった複雑性の特徴がある [1]．これに対応可能な手法は存在するが，どちらか一方の複雑性にしか対応することができないため，ノイズとオーバーラップが両方生じているデータセットには適用が困難である．

そこで本研究では，複雑性を考慮しつつデータの不均衡性の解消を目指す．

2 関連研究

データの複雑性を定量化する研究として，k-Disagreeing Neighbors（kDN）と Linear Discriminant Analysis（LDA）が挙げられる．kDN はノイズを定量化する手法であり，k 近傍法（kNN）を用いて学習によって誤判別する要素を定量化することができる．また，LDA はオーバーラップを定量化する手法であり，線形判別を用いて学習によって誤判別する要素を定量化することができる．

データの複雑性に対応可能なリサンプリング手法として，SMOTE-ENN[2]，SMOTE-Tomek[3] が挙げられる．オーバーサンプリングとして一般的な SMOTE によって少数派を増やした後，SMOTE-ENN[2] では，誤分類された要素を取り除く Edited Nearest Neighbors（ENN）によってノイズに対応可能なアンダーサンプリングを，SMOTE-Tomek[3] では，データ間の距離を考慮できる Tomek Links によってオーバーラップに対応可能なアンダーサンプリングを行う．

しかし，これらの手法ではどちらか一方の複雑性にしか対応することができないため，ノイズとオーバーラップが両方生じているデータセットには適用が困難である．また，使用するデータにおいて複雑性の存在を仮定しているため，実際には生じていない場合にも要素を取り除いてしまう可能性がある．

3 提案手法

本研究では，データ内のノイズとオーバーラップに対応しつつ不均衡性の解消を目指し，事前に計算した二種類の複雑性に基づくリサンプリング手法を提案する．リサンプリングにはノイズとオーバーラップに対応が可能な ENN と Tomek Links を用い，不均衡性に加え複雑性にも対応する．しかし，二つのアンダーサンプリングを同時に扱う場合，生じている複雑性の比率を考慮しなければ，要素を過度に取り除いてしまい分類精度を低下させてしまう可能性がある．

このため，各リサンプリング手法の比率と順序を決定する必要がある．リサンプリングの比率は，ノイズとオーバーラップの値を事前に計算し，ENN の比率を式 (1) の $r_n(\mathcal{D})$ と，Tomek Links の比率を式 (2) の $r_o(\mathcal{D})$ という二つの比率決定関数を用いて決定する．

$$r_n(\mathcal{D}) = \begin{cases} \frac{0.5LDA_o(\mathcal{D})}{kDN_n(\mathcal{D})} & \text{if } kDN_n(\mathcal{D}) = LDA_o(\mathcal{D}) \\ \\ \frac{kDN_n(\mathcal{D})}{kDN_n(\mathcal{D})+LDA_o(\mathcal{D})} & otherwise \end{cases} \tag{1}$$

$$r_o(\mathcal{D}) = 1 - r_n(\mathcal{D}) \tag{2}$$

計算したノイズ $kDN_n(\mathcal{D})$ とオーバーラップ $LDA_o(\mathcal{D})$ が等しい場合は，$r_n(\mathcal{D})$ と $r_o(\mathcal{D})$ の和が 1 になるように $r_n(\mathcal{D})$ は 0.5 を返し，等しくない場合は，ノイズとオーバーラップの合計に対するノイズの比率を返す．

ENN は，近傍の要素から誤分類された要素の特定を行うが，クラス同士の境界に位置する要素のような，他の要素から離れている要素は取り除くことができない．一方 Tomek Links は，要素同士の距離を考慮して要素を取り除くため，別のクラスに含まれる要素を取り除くことはできない．リサンプリングの順序を誤ると，各手法が対応していない複雑性に対してアンダーサンプリングが行われ，取り除く必要のない要素を取り除いてしまう可能性がある．そのため，リサンプリングの順序は，ノイズとオーバーラップの大小によって式 (3) の順序決定関数 $f(\mathcal{D})$ を用いて決定する必要がある．式 (3) は条件を満たす場合は 1 を，満たさない場合は 0 を

A Resampling Method for Imbalanced Datasets Considerating Data Complexity
†Taisho Sasada ‡Tokiya Baba ‡Yusuke Kimura †Kenji Hatano
†Faculty of Culture and Information Science, Doshisha University
‡Graduate School of Culture and Information Science, Doshisha University

表 1: 評価実験の結果

			正解率							F 値						
ID	多数派	少数派	既存手法	cnf	cyf	dnf	dyf	inf	iyf	既存手法	cnf	cyf	dnf	dyf	inf	iyf
D1	689	42	0.81	0.79	0.81	0.80	**0.83**	0.79	0.82	0.72	0.66	0.70	0.68	**0.75**	0.67	0.72
D2	1,268	673	0.94	0.76	**0.95**	0.94	**0.95**	0.84	**0.95**	0.91	0.86	**0.95**	0.90	0.94	0.88	0.94
D3	357	212	0.96	0.95	0.97	0.96	**0.98**	0.95	0.97	0.94	0.90	0.92	0.93	**0.97**	0.91	0.94
D4	7,963	2,037	0.88	0.86	**0.90**	0.87	0.89	0.86	0.89	0.88	0.85	**0.90**	0.86	0.87	0.85	0.88
D5	290	10	0.91	0.89	0.92	0.90	**0.94**	0.89	0.93	0.90	0.88	0.89	0.90	**0.93**	0.89	0.91
D6	224	81	0.92	0.87	0.91	0.91	**0.95**	0.89	0.93	0.90	0.87	0.89	0.89	**0.93**	0.88	0.91
D7	16,259	1,639	**0.98**	0.97	**0.98**	**0.98**	**0.98**	0.97	**0.98**	**0.97**	0.95	**0.97**	0.96	0.96	0.95	**0.97**
D8	10,422	1,908	**0.97**	0.94	0.96	0.95	0.95	0.94	0.95	**0.96**	0.89	0.94	0.92	0.93	0.90	0.93
D9	4,913	329	**0.98**	0.96	0.97	0.97	0.97	0.96	0.97	**0.95**	0.92	0.93	0.94	0.94	0.93	0.93
D10	4,898	1,599	0.97	0.96	**0.98**	0.97	**0.98**	0.96	**0.98**	0.96	0.95	0.96	0.96	**0.98**	0.95	**0.98**

出力する．1 が返された場合はノイズの方が大きいため ENN を，0 が返された場合はオーバーラップの方が大きいため Tomek Links を先に行う．

$$f(\mathcal{D}) = \begin{cases} 1 & \text{if } kDN_n(\mathcal{D}) \geq LDA_o(\mathcal{D}) \\ 0 & \text{if } kDN_n(\mathcal{D}) < LDA_o(\mathcal{D}) \end{cases} \quad (3)$$

複雑性の定量化には，誤判別の要因となる要素の数を使用するが，データセットの大きさによって異なるため，データセットの大きさによる影響を考慮できていない．したがって，様々な不均衡データで複雑性が計算可能な形に標準化する必要がある．

したがって本研究では，クラス単位とデータセット単位による標準化は次のようにして行う．クラス単位の複雑性は誤判別の要因となる要素の数を各クラスにおける要素数で割り，二つの比率の調和平均を計算することでクラスにおける複雑性を算出する．また，データセット単位の複雑性は誤判別の要因となる要素の数をデータセット全体の要素数で割って算出する．クラスとデータセット双方の複雑性を表すために，クラスにおける複雑性とデータセットにおける複雑性の調和平均を計算する．これは双方の複雑性を考慮するために適合率と再現率，F 値の関係に倣う．

4 評価実験

提案手法の有用性を検証するため，決定木，サポートベクターマシン，ロジスティック回帰を使用し，UCI Machine Learning Repository*の不均衡データセット 10 種類を用いて層化 10 分割交差検証を行う．既存のリサンプリング手法と提案手法でそれぞれ最も正解率，F 値の高かった結果を表 1 に示す．表 1 は既存手法と，提案手法である cnf（クラス単位の複雑性+$f(\mathcal{D})$ なし），cyf（クラス単位の複雑性+$f(\mathcal{D})$ あり），dnf（データセット単位の複雑性+$f(\mathcal{D})$ なし），dyf（データセット単位の複雑性+$f(\mathcal{D})$ あり），inf（複雑性の統合指標+$f(\mathcal{D})$ なし），iyf（複雑性の統合指標+$f(\mathcal{D})$ あり）を列とし，データセットの ID を行としている．なお，各データセットにおいて最も高い値を示したものを太字で示す．

実験の結果，D1〜D6，D10 のようなノイズとオーバーラップの値に絶対差が生じているデータセットでは，リサンプリングの比率や順序を活用することで正解率と F 値が向上した．しかし，D7〜D9 のような計算したノイズとオーバーラップの値に絶対差が生じていないデータセットでは，正解率は向上しなかった．これはノイズとオーバーラップの間に絶対差が生じていなかったため，比率や順序を変更しても差がなかったと考えられる．

5 おわりに

本稿では，データの複雑性を考慮するために複雑性を事前計算し，計算結果に基づくリサンプリングを行った．提案手法は既存手法よりも頑健なリサンプリングを行うことが可能であり，正解率と F 値において向上することが確認された．

今後の課題として，よりデータの複雑性を緻密に捉えた形で指標の統合を行う必要がある．

謝辞

本研究の一部は JSPS 科研費 JP18H03342 と JP19H01138 の助成を受けたものである．

参考文献

[1] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. *Information sciences*, Vol. 250, pp. 113–141, 2013.

[2] Gustavo Enrique Batista, Ana Lucia Cetertich Bazzan, and Maria Carolina Monard. Balancing Training Data for Automated Annotation of Keywords: a Case Study. *Journal of artificial intelligence research*, Vol. 3, No. 2, pp. 15–20, 2003.

[3] Gustavo Enrique Batista, Ronaldo Cristiano Prati, and Maria Carolina Monard. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explorations*, Vol. 6, No. 1, pp. 20–29, 2004.

*UCI Machine Learning Repository.
http://archive.ics.uci.edu/ml（閲覧日 2020 年 1 月 10 日）