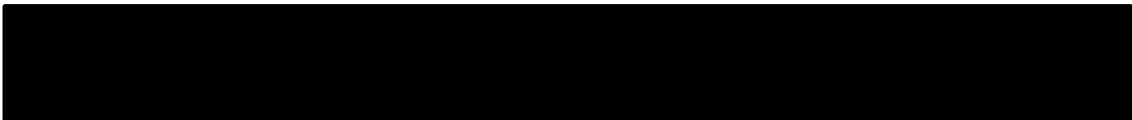
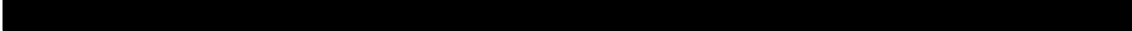


研究助成金交付申請書

公益財団法人 栢森情報科学振興財団

理事長 栢 森 雅 勝 殿

2019年 8月 23日

フリガナ コマミズ タカヒロ 〈申請者氏名〉 駒水 孝裕		〈年齢〉 32	
フリガナ 〈所属機関名 (学部・学科・役職) 省略せずにご記入ください〉 名古屋大学情報基盤センター 助教			
〈ADD〉 〒464-8601 愛知県名古屋市千種区不老町 名古屋大学 情報基盤センター			
〈TEL〉 052-789-4359		〈FAX〉	
〈E-Mail〉 taka-coma@acm.org			
自宅 〈ADD〉 			
〈TEL〉 			
〈学歴・職歴 (大学入学から) 〉 2005年 4月 筑波大学第三学群情報学類 入学 2009年 3月 筑波大学第三学群情報学類 卒業 2009年 4月 筑波大学システム情報工学研究科コンピュータサイエンス専攻 博士前期課程 入学 2011年 3月 筑波大学システム情報工学研究科コンピュータサイエンス専攻 博士前期課程 修了 2011年 4月 筑波大学システム情報工学研究科コンピュータサイエンス専攻 博士後期課程 入学 2015年 3月 筑波大学システム情報工学研究科コンピュータサイエンス専攻 博士後期課程 修了 2015年 4月 筑波大学計算科学研究センター 研究員 理研・文科省プロジェクト「実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発」に従事 2018年 2月 名古屋大学情報基盤センター 助教 現在に至る			
〈助成金申請額〉 200万円			
その用途 (内訳)			
内 容	金 額(千円)	内 容	金 額(千円)
物品費		旅費	
● 実装・実験用のワークステーション	500	● 国内研究会参加・発表 (7万円×2回)	140
● 資料収集・成果発表用のノートPC	300	● 国内シンポジウム参加・発表 (11万円×2回)	220
● 消耗品	20	● 国際会議 (アメリカ) 参加・発表 (40万円×2回)	800
		その他	
		● 事務処理経費 (1%)	20

<研究テーマ> 不均衡データに対する異なる比率のサンプリングを利用した協調学習

<研究の概要、研究期間と学術的意義>

【研究概要】

不均衡データはラベル付きデータにおいてラベルごとのサンプル数が偏ったデータである。不均衡データは分類問題において分類性能を低下させる要因の一つである。不均衡データが問題になる例として、クレジットカード不履行予測やネットワーク侵入検知、電子商取引における商品推薦などがある。本研究では、不均衡データへの対処として有効とされるリサンプリングに注目する。リサンプリングには、多数派データをサンプリングするアンダーサンプリング (US) と少数派データをかさ増しするオーバーサンプリング (OS) の二種類がある。本研究では US に着目し、サンプリング比率を変化させることで分類器が捉えるデータ特性が変わることを利用した分類手法の構築を目指す。リサンプリングに関する研究動向については、資料2に付記する。

具体的には、複数のサンプリング比率における US を組合せた分類器を構築する。US において異なるサンプリング比率を用いると多数派あるいは少数派に寄った分類器を構築可能である(資料1を参照)。多数派(少数派)に寄った分類器は、多数派(少数派)の分類を得意とし、多数派(少数派)を高い確度で分類する。この確度をデータのラベルへの信頼度合いと捉え、様々な比率による分類器から信頼度の高い分類を実現する。

研究提案者はこれまでの研究 [1,2] (資料1) にて、上記基本アイデアを実現するためのフレームワークの構築を行った。本研究では、フレームワークにおいてより深化すべき点について掘り進める。具体的には、サンプリング比率の導出方法、サンプリング比率に対する重要度の算出方法の二点を洗練する。また、実社会における不均衡データに対する大規模なデータセットに対して本研究を適用し、本研究の有効性および不得手なデータセットの特徴について解析する。データの不均衡性は分類性能低下の唯一の要因ではないため、他の要因について上記解析で明らかにする。

【研究期間：二年間】

●一年目：複数のサンプリングレートにおける US を組合せた分類器の洗練。

提案者がすでに進めている研究 [1,2] (資料1) について実験・改良を重ね分類器を構築する。特に、サンプリング比率の導出方法、サンプリング比率に対する重要度の算出方法について研究する。

●二年目：大規模データセットを用いた分類実験・解析に基づく手法改良。

提案手法の得手・不得手を明らかにするとともに、不均衡性以外の問題点を明らかにする。解析結果をもとに、提案手法のさらなる改善を目指す。

【学術的意義】

不均衡データへの対処は盛んに研究されている。リサンプリングはその中でも様々なデータへの応用が可能な手法である。リサンプリングの研究において、異なる分布やサンプリングレートを組合せる手法は未だに研究されておらず、不均衡データ問題に対する新たな視点を与えることが期待される。

<研究業績一覧(著者名・論文テーマ・学会誌名・巻・号・出版年月・ページ数など)>

※申請研究に関連のある主なもの※

- [1] 植原 リサ, 駒水 孝裕, 小川 泰弘, 外山 勝彦, 弱分類器の調整に基づく不均衡データ向けアンサンブル・フレームワーク, 第12回 Web とデータベースに関するフォーラム, 2019年(採択済み)
- [2] Yasuhiro Ogawa, Michiaki Satou, Takahiro Komamizu, Katsuhiko Toyama, "nagoy Team's Summarization System at the NTCIR-14 QA-Lab PoliInfo", Proc. NTCIR-14, pp.182-189, 2019
- [3] Takahiro Yamakoshi, Takahiro Komamizu, Yasuhiro Ogawa, Katsuhiko Toyama, "Japanese Legal Term Correction using Random Forests", in Proc. JURIX 2018, pp.161-170, 2018 (Best paper award)
- [4] Takahiro Komamizu, "Learning Interpretable Entity Representation in Linked Data", in Proc. DEXA 2018, pp.153-168, 2018
- [5] Takahiro Komamizu, Yasuhiro Hayase, Toshiyuki Amagasa, Hiroyuki Kitagawa, "Exploring Identical Users on GitHub and Stack Overflow", in Proc. SEKE 2017, pp.584-589, 2017
- [6] Takahiro Komamizu, Toshiyuki Amagasa, Hiroyuki Kitagawa, "SPOOL: A SPARQL-based ETL Framework for OLAP over Linked Data", in Proc. iiWAS 2015, pp.49:1-10, 2015 (Best paper award)

1. 英文で記入する場合は、和文も必ず併記してください。
2. 資料を添付する場合は、A4用紙2枚までとします。

[1] では、本研究提案の基盤となるフレームワーク (MUE) を提案した。

MUE における基本アイデアは図 1 に示す予備実験の結果から得た。予備実験において、US に異なるサンプリング比率を用いた場合、分類性能の性質が異なることを示した。図 1 はサンプリング比率を 0.1 から 1 と 1 から 10 に変化させたときの分類性能を示している。ここでは分類性能として、真陽性率 (少数派における再現率, True Positive Rate ; 青) と真陰性率 (多数派における再現率, True Negative Rate ; 赤) を積み上げグラフで表示している。灰色の部分には分類に失敗した割合を表す。図 1 から、サンプリング比率を 1 以下にした場合、少数派データに対する分類性能が向上する一方で多数派データに対する分類性能は低下し、サンプリング比率を 1 以上にした場合、逆のことが起こることを確認した。

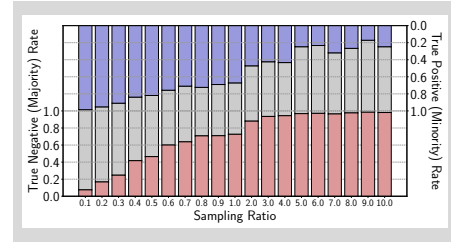


図 1

MUE の狙いは、図 1 に示した異なるサンプリング比率における分類器の性質の違いをアンサンブル学習によって吸収し、より分類性能の高い分類器を構築することである。この実現のために、(1) 異なるサンプリング比率で学習した分類器を組み合わせる枠組みの構築 (図 2), (2) 導入するサンプリング比率の選択, (3) サンプリング比率ごとに学習される分類器の重み付け, を行った。MUE では、まず、決められたサンプリング比率に基づきアンダーサンプリングを適用し、学習用データを生成する (Sampling Phase)。

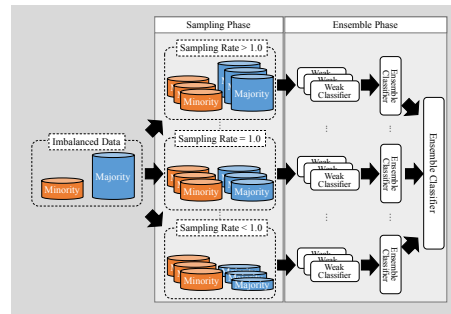


図 2

次に、Ensemble Phase において、それぞれの学習データに対して弱分類器を学習する。サンプリング比率ごとに学習した弱分類器をアンサンブル学習によりひとつの分類器としてまとめる。最後に、各サンプリング比率で学習した分類器をアンサンブル学習により最終的な分類器を構築する。

表 1 に提案フレームワーク (MUE) と既存手法と比較した結果を示す。比較手法は OS 手法三種類と US 手法三種類 (資料 2 を参照) で、データセットは、UCI 機械学習リポジトリと Kaggle dataset から取得した不均衡性を有する実データを用いた。評価指標には、真陽性率と真陰性率の幾何平均を用いた。この表から、US の方が OS よりも良い分類性能を示している事がわかる。特に、MUE は既存手法に対して良い性能を示している。MUE が既存手法で最高の性能を示した EasyEnsemble (EE) に対して劣っているデータにおいてその差は小さい。注目したいのが、少数派データ数 (#minor) が小さいデータにおける MUE の性能である。少数派データ数が 10^2 を下回る場合に、MUE はより高性能であることがわかる。少数派データ数が多い場合でも MUE は既存手法に対し、同等かそれ以上の性能を示している。このことから、複数のサンプリング比率を導入することの有効性が示された。

表 1

Dataset	#minor	ORG	Oversampling			Undersampling			MUE
			SMT	ADA	SWIM	RUS	RBST	EE	
D10	20	.511	.283	.440	.686	.570	.565	.629	.779
D1	42	.494	.577	.550	.635	.682	.560	.714	.750
D9	63	.470	.581	.538	.587	.639	.348	.700	.685
D2	68	.885	.918	.902	.969	.915	.966	.940	.967
D14	136	.496	.542	.595	.557	.758	.645	.861	.862
D12	492	.872	.871	.844	.911	.901	.890	.937	.938
D17	678	.707	.754	.733	.776	.781	.685	.813	.811
D13	803	.807	.838	.833	.831	.881	.818	.913	.913
D5	1,639	.904	.918	.913	.921	.909	.891	.935	.938
D8	1,813	.903	.910	.900	.899	.898	.933	.946	.940
D6	1,908	.722	.744	.744	.713	.791	.703	.845	.845
D11	2,037	.640	.663	.653	.653	.670	.634	.756	.753
D7	2,091	.834	.830	.831	.834	.855	.712	.926	.924
D16	2,190	.755	.771	.758	.768	.787	.664	.757	.761
D19	3,690	.805	.791	.755	.805	.791	.778	.834	.830
D4	6,636	.587	.583	.581	.577	.620	.536	.701	.705
D3	9,493	.885	.919	.916	.885	.928	.859	.858	.831
D18	31,877	.675	.689	.689	.675	.712	.617	.779	.780
D15	50,611	.467	.471	.472	.468	.539	.423	.611	.607
Macro Avg.	< 10^2	.590	.590	.608	.719	.702	.610	.746	.795
	< 10^3	.683	.698	.705	.764	.782	.688	.827	.838
	< 10^4	.722	.735	.734	.765	.787	.717	.827	.837
	all	.706	.719	.718	.745	.770	.696	.813	.822

本研究提案では、MUE におけるサンプリング比率の導出方法とサンプリング比率に対する重要度の算出方法について研究することでさらなる性能改善を目指す。また、表 1 から MUE を含む US 手法が OS 手法に劣るデータが存在する (D2, D3, D16)。同様に、MUE が EE に明らかに劣るデータも存在する (D9)。これらはデータの不均衡性以外の性質によるものと推測される。これらのデータと他のデータの性質の違いを深く分析することで MUE をより高性能にすることが期待される。

(資料2) リサンプリングに関する研究動向

まず、不均衡データの問題について説明する。図3にデータの例を示す。青色の丸が多数派の観測データを表し、橙色の三角が少数派の観測データを表す。図中のグラデーションはそれぞれのデータの真の分布を表現する。観測データに基づく分類器では、分類の境界付近のデータは多数派に分類される可能性が高い。図中の赤い点Xは少数派データの分布に属しているが、近傍のデータの数から多数派に分類させてしまう可能性が高い。このような観測データ（特に少数派データ）の分布の偏りにより、分類器の性能を低下させてしまう。

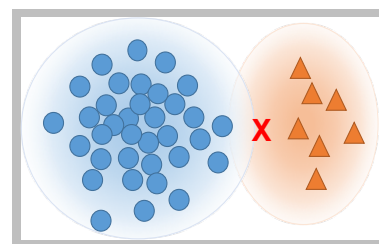


図3

この問題に対処するのが、リサンプリングである。リサンプリングには、多数派データを減らすアンダーサンプリング (US) と少数派データを増やすオーバーサンプリング (OS) がある。US は図中の青丸のデータを間引いて、多数派データの数を少数派データの数に合わせる (資料1でのRUS)。US はサンプリングした際に多数派データを破棄してしまう。これでは重要なデータを見落とすこともあるため、多数派データももれなく使う方が良くとされている。これを実現したのが EasyEnsemble [1] である。EasyEnsemble は多数派データをカバーする回数だけ US を行う。サンプリングされたデータそれぞれについて弱分類器を学習し、弱分類器を統合するアンサンブル学習を行うことで上記問題を解消した。別なアプローチとして、少数派データと分類が困難な多数派データをサンプリングする手法も提案されている。これは、多数派と少数派の境界にあるデータをサンプリングすることでより分類性能の高い分類器の構築を目指すものである。このアプローチを実現した手法として、RUSBoost [2] (資料1でのRBST) や Trainable US [3] である。[3] については、資料1では実験に含めていない。これは、[3] で示した有効性が正解率をもとにしており、少数派データがうまく分類できなくとも高いスコアになる評価指標を使っていたため、同じ実験データ・評価指標で実験したところ性能が著しく悪かったために除外した。

一方で、少数派データを (単純に) 複製する OS は過学習を起こすため、「少数派データのような」データを生成する Synthetic Oversampling (SO) が研究されている。SO の代表的な手法である SMOTE [4] (資料1でのSMT) は少数派データの観測データの近傍に疑似サンプルを生成する。つまり、図3中の三角のデータを付近に疑似サンプルを生成する。これにより、数の偏りによる少数派データの軽視は避ける。SMOTE の拡張手法は多く提案されている。ADASYN [5] (資料1でのADA) もその一つである。しかしながら、疑似サンプルの分布が少数派データの観測データに依存するため、観測データが偏っている場合には、真の分布に近づけることは難しい。これに対して、SWIM [6] では、少数派データの分布の広がりが多い多数派データの分布の広がりと同じ広がりを持つ疑似サンプルを生成する。つまり、図中の丸のデータの広がりと同じ広がりを持つ疑似サンプルを生成する。これにより、より広い範囲の疑似サンプルを生成できる。

[参考文献]

- [1] X. Liu et al., Exploratory Undersampling for Class-Imbalance Learning. IEEE Trans. Systems, Man, and Cybernetics, Part B, 39(2):539–550, 2009.
- [2] C. Seiffert et al., RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. IEEE Trans. Systems, Man, and Cybernetics, Part A 40(1):185–197, 2010
- [3] M. Peng, Trainable Undersampling for Class-Imbalance Learning. In AAI 2019, pp. 4707–4714, 2019
- [4] N. V. Chawla et al., SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res., 16:321–357, 2002.
- [5] H. He et al., ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In IJCNN 2008, 1322–1328, 2008
- [6] S. Sharma et al., Synthetic Oversampling with the Majority Class: A New Perspective on Handling Extreme Imbalance. In ICDM 2018, pages 447–456, 2018.

(資料3) 助成金の使途に関する補足

- 物品費は、研究をより効率的に進めるための計算資源の拡充のために用いる。また、必要に応じて消耗品の補充も行う。
- 旅費は、研究成果の発表および研究内容について議論するために用いる。研究内容について国内で議論し深化するために、国内の研究会 (WebDB Forum を想定) と国内シンポジウム (DEIM を想定) への参加を計画している。また、研究成果を国際的に発表するために、国際会議 (AAAI を想定) に参加・発表を計画している。それぞれ、研究期間の一年目と二年目に一回ずつ参加する予定である。
- その他は、所属部署にて研究助成金を管理するのに必要な費用 (助成金の1%) である。