CF-009

不均衡マルチクラスデータのためのクラス分類確率を特徴量として用いる半教師あり学習
Semi-supervised Learning Using Class Probabilities as Features for Imbalanced Multi-class Data

吉田 由起子† 竹林 知善†
Yukiko Yoshida Tomoyoshi Takebayashi

1. はじめに

機械学習モデルの構築には、学習データとして多量の正解ラベルありデータが必要である。しかし、入力データへの正解ラベルの付与には多大の時間とコストがかかり、少量の正解ラベルありデータしか用意できないことがあり、モデル構築を困難にしている。一方で、入力データはセンサー等によって自動的に多量に取得できる場合が多い。そこで、多量のラベルなしデータと少量のラベルありデータの情報を組み合わせることによってラベルなしデータに疑似ラベルを付与し、モデル構築に利用するという半教師あり学習 (semi-supervised learning (SSL)) アプローチが注目されている。しかし、クラス間の出現頻度に大きい偏りがある場合、SSL が低頻度クラスの疑似ラベルを生成しにくいという問題がある。その問題への対処方法としては、教師あり学習で用いられている手法と同様に、誤分類コストの重みを調整する方法や, 高頻度クラスのデータ削減や低頻度クラスのデータ合成といったサンプリング調整手法 (under-sampling, over-sampling, SMOTE[1]) が知られているが、学習データとテストデータでのクラス間頻度に相違があると誤分類コストの調整が役に立たない可能性があり、サンプリング調整法ではクラス分類にとって重要なデータを削減したり、誤ったデータを合成してしまうおそれがある。とくにマルチクラスの場合には適切なバランスの取り方が難しい。この問題を解決するアプローチとして、ラベルありデータセットに対して所定のクラス分類器を用いた分割交差検証によって得られる各データ点のクラス別分類確率に着目した SSL の疑似ラベル付与方法を考案した。本稿では、提案手法の処理手順を説明し、不均衡マルチクラスデータセットへの適用実験による従来法との性能比較について報告する。

2. 半教師あり学習について

一般的な SSL の方法は、特徴量空間におけるラベルありデータとラベルなしデータの分布およびラベルありデータのラベル情報に基づいて、各ラベルなしデータ点に対して各クラスの分類確率を算出し、分類確率が最も高いクラスをそのデータ点の疑似ラベルとして選択する。ここで、特徴量空間上でいくつかのクラスが高頻度で広く分布し、それとは別の非常に低頻度のクラスが局在的に分布しているような不均衡マルチクラスデータセットを考える。たとえば4つのクラス c_1, c_2, c_3, c_4 が存在し、出現頻度は $c_1 > c_2 > c_3 > c_4$ のような高低差があり、とくに c_4 が非常に低頻度であるとする。各ラベルなしデータ点について各クラス c_i の分類確率 p_i $(\mathrm{i} = 1, \cdots, 4)$ を計算する際には、低頻度クラス c_4 よりも他の高頻度のクラスの影響をより大きく受けるため、p_4 は p_i $(\mathrm{i} = 1, \cdots, 4)$ の中で最大値になりにくい。しかしながら、各データ点のクラス別分類確率のベクトル (p_1, p_2, p_3, p_4) は、そのデータ点周辺でどのクラスがどのように分布しているかを反映したパターンを持つものと考えられる。

† (株) 富士通研究所, Fujitsu Laboratories, Ltd.

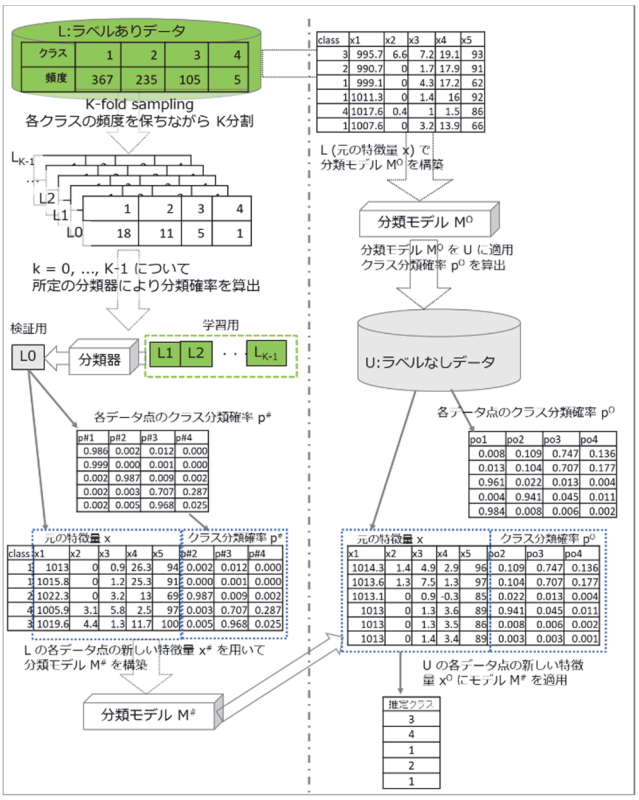

図 1. 提案手法

3. 提案手法

我々は特徴量空間上のデータ点のクラスの分布とクラス別分類確率のパターンとの関係性に注目し、クラスが未知のデータ点は特徴量空間上で近くに分布するクラス分類確率のパターンが似たデータ点と同じクラスに属する可能性が高いという仮説を立てる。そして、この仮説に基づいて、非常に低頻度のクラスについてもラベルなしデータへの疑似ラベル付与を可能とする SSL 手法を考案した。

その方法は、まず、所定のクラス分類器によってラベルありデータに対して分割交差検証を行い、その結果得られる各ラベルありデータ点に対するクラス別分類確率ベクトル $\mathrm{p}^{\#}$ を、元の特徴量ベクトル x と組み合わせて新しい特徴量ベクトル $\mathrm{x}^{\#}$ を構成し、それらを用いてクラス分類モデル $\mathrm{M}^{\#}$ を構築する。つぎに、ラベルありデータの元の特徴量ベクトルを用いて構築しておいたクラス分類モデル M^{O} を各ラベルなしデータ点に対して適用してクラス別分

類確率 p^{O} を算出し、それらを元の特徴量ベクトルと組み合わせて上記と同様の特徴量ベクトル x^{O} を構成し、$\mathrm{M}^{\#}$ を適用することによりクラスを推定するというものである。

その処理の流れを下記に示す:

1. ラベルありデータセット L を K 個の部分データセット $\mathcal{L}=\{L_0,\cdots,L_{K-1}\}$ に分割する。ここで、K は分割された部分データセットのサイズが L に対して十分に小さくなるように設定する。
2. 各 $L_k \in \mathcal{L}$ $(\mathrm{k}=0,1,\cdots,\mathrm{K}-1)$ について、つぎの 3, 4 を処理する:
3. L_k を検証用データセットとし、L から L_k を除いたデータセット $\mathrm{L}\setminus L_k$ を学習データとして所定のクラス分類器により、L_k の各データ点 w についての各クラス c_i $(\mathrm{i}=1,\cdots,\mathrm{N})$ の分類確率 $p_i^{\#}$ を算出する。
4. L_k の各データ点について元の特徴量 $(x_1,\cdots,x_J)$ とクラス別分類確率ベクトル $\left(p_1^{\#},\cdots,p_N^{\#}\right)$ を組み合わせて、新しい特徴量ベクトル $\mathrm{x}^{\#}=(x_1,\cdots,x_J,p_2^{\#},\cdots,p_N^{\#})$ を構成する。ここで $p_1^{\#}+\cdots+p_N^{\#}=1$ であることから、$\left(p_1^{\#},\cdots,p_N^{\#}\right)$ から 1 変数 $p_1^{\#}$ を除いて $\mathrm{x}^{\#}$ に組み入れるものとする。
5. すべての $L_k \in \mathcal{L}$ のデータ点についての新しい特徴量ベクトル $\mathrm{x}^{\#}$ を用いてクラス分類モデル $\mathrm{M}^{\#}$ を構築する。
6. ラベルあり L (元の特徴量ベクトル) データを用いて 3 と同じクラス分類器によりラベルなしデータ U の各データ点のクラス別分類確率 p_i^{O} $(\mathrm{i}=1,\cdots,\mathrm{N})$ を算出し、4 と同様にして元の特徴量 $(x_1,\cdots,x_J)$ と組み合わせて新しい特徴量ベクトル $\mathrm{x}^{O}=(x_1,\cdots,x_J,p_2^{O},\cdots,p_N^{O})$ を構成する。
7. U の各ラベルなしデータ点 v についての新しい特徴量ベクトル x^{O} にクラス分類モデル $\mathrm{M}^{\#}$ を適用して推定されたクラスを v の疑似ラベルとする。

本提案手法では、ラベルありデータを余分に増やしたり減らしたりせずに均等に使用するので、正解クラスの分布の性質を崩さずにラベルなしデータに反映させることができる。また、学習データとテストデータでクラス間の出現頻度が異なる場合にとくに問題となる誤分類コストのクラスごとの重みの調整を行わずにすむという利点がある。

4. 実験

4.1 データセット概要

表 1 実験用データセット情報

使用データ	日本気象 (株) 気象予報配信データ「今日明日天気（市区町村）」[3]				
分析対象	北関東 3 都市: 水戸市, 宇都宮市, 前橋市 南関東 4 都市: さいたま市, 千葉市, 千代田区, 横浜市				
データ期間	学習用: 2012/05/01～2017/04/30 テスト用: 2017/05/01～2018/04/30				
入力	各日の朝 6 時の気象予報データ (気温、相対湿度、大気圧、降水量、風速、風向、雲量)				
出力	その日の朝 6 時の天気予報値 (1.晴/2.曇/3.雨/4.雪の 4 クラス)				
クラス	1.晴	2.曇	3.雨	4.雪	

表 2 SSL 実験評価用データセット情報

クラス別頻度	1.晴	2.曇	3.雨	4.雪	総数
Ls	367 (0.515)	235 (0.330)	105 (0.147)	5 (0.007)	712
Us	3105 (0.485)	2175 (0.339)	1085 (0.169)	43 (0.007)	6408
Un	2718 (0.449)	1719 (0.284)	830 (0.137)	73 (0.012)	5340
Ts	624 (0.483)	443 (0.343)	216 (0.167)	9 (0.007)	1292
Tn	514 (0.530)	286 (0.295)	164 (0.169)	5 (0.005)	969

本手法を、関東 7 都市の各日の気象予報データを用いてその日の天気 (晴、曇、雨、雪の 4 クラス) を推定するクラス分類問題に適用した。表 1 に実験で用いたデータセットの情報を示す。このデータセットでは 晴 > 曇 > 雨 > 雪の順で頻度が高く、とくに雪の頻度が非常に低いという特徴がある。2012/05/01～2017/04/30 の 5 年分を学習用データセットとして用い、2017/05/01～2018/04/30 の 1 年分をテスト用データセットとして用いる。つぎに、SSL 実験評価用に、学習用データセットをラベルありデータセットとラベルなしデータでセットに振り分ける。ここで、ラベルありデータとラベルなしデータとの間で特徴量の分布やクラスの出現頻度に違いがある場合の SSL の性能を評価するために、データセットを北関東 3 都市と南関東 4 都市に分けて扱うこととし、データセットを下記のように分割した:

- ラベルあり L_S: 南関東 4 都市の学習用データセットからランダムに 10% を選択したもの
- ラベルなし U_S: 南関東 4 都市の学習用データセットの残りの 90%
- ラベルなし U_N: 北関東 3 都市の学習用データセット
- テスト用 T_S: 南関東 4 都市のテスト用データセット
- テスト用 T_N: 北関東 3 都市のテスト用データセット

表 2 に SSL 実験評価用データセットの情報を示す。

4.2 ラベルあり・ラベルなし・テストデータセットの種々の組み合わせの下での従来法と提案手法の比較実験

これらデータセットに対して、SSL を用いない従来法 (ラベルありデータのみでクラス分類モデルを構築) と SSL を用いる従来法と提案手法を適用してテストデータセットのクラス分類性能を比較する。SSL のベース手法として self-training を用い、分類器として LIBSVM[4] の SVM one-vs-one を使用する。self-training を 1 回実行するごとに、各ラベルなしデータ点のクラス別分類確率を算出し、0.9 以上のクラス分類確率を持つデータ点に対して当該クラスを疑似ラベルとして付与する。疑似ラベルを付与されたデータ点をラベルなしデータセットからラベルありデータセットに移動することにより、ラベルありデータセットとラベルなしデータセットを更新する。すべてのラベルなしデータ点に疑似ラベルが付与されるか、0.9 以上のクラス分類確率を持つラベルなしデータ点が見つからなくなるまで

self-training を繰り返す。提案手法の K-分割交差検証のパラメーターは K = 20 とする。
今回、ラベルあり・ラベルなし・テストデータセットを様々に組み合わせて、従来法と提案手法による適用実験を実施し、テストデータに対するクラス推定精度を比較した。実験で使用した手法とデータセットの組み合わせを下記の形式で表記することとする:

比較手法の略称: ラベルありデータ (+ ラベルなしデータ + ...) → テストデータ

ここで、比較手法の略称は下記のとおりである:

- non-SSL: SSL を用いない従来法
- SSL: 従来の SSL
- CP-SSL: 提案手法 (クラス別分類確率 Class Probabilities を特徴量化)

たとえば、「non-SSL: $L_S \rightarrow T_S$」は、SSL を用いない従来法でラベルありデータ L_S を用いてモデルを構築し、テストデータ T_S に適用した実験、「SSL: $L_S + U_S \rightarrow T_S$」は、従来の SSL でラベルありデータ L_S とラベルなしデータ U_S を組み合わせて疑似ラベルを生成してモデルを構築し、テストデータ T_S に適用した実験、「CP-SSL: $L_S + U_S + U_N \rightarrow T_N$」は、提案手法でラベルありデータ L_S とラベルなしデータ U_S と U_N を組み合わせて疑似ラベルを生成してモデルを構築し、テストデータ T_N に適用した実験を表す。

(従来法) non-SSL: $L_S \rightarrow T_S$

従来法		推定クラス				再現率
		1.晴	2.曇	3.雨	4.雪	
正解クラス	1.晴	623	0	1	0	0.998
	2.曇	25	357	61	0	0.806
	3.雨	2	42	172	0	0.796
	4.雪	0	3	6	0	0.000
適合率		0.958	0.888	0.717	-	0.892

(従来法) non-SSL: $L_S \rightarrow T_N$

従来法		推定クラス				再現率
		1.晴	2.曇	3.雨	4.雪	
正解クラス	1.晴	514	0	0	0	1.000
	2.曇	8	275	3	0	0.962
	3.雨	1	85	78	0	0.476
	4.雪	0	5	0	0	0.000
適合率		0.983	0.753	0.963	-	0.895

(従来法) SSL: $L_S + U_S \rightarrow T_S$

従来法		推定クラス				再現率
		1.晴	2.曇	3.雨	4.雪	
正解クラス	1.晴	624	0	0	0	1.000
	2.曇	25	339	79	0	0.765
	3.雨	2	58	156	0	0.722
	4.雪	0	4	5	0	0.000
適合率		0.958	0.845	0.650	-	0.866

(従来法) SSL: $L_S + U_S \rightarrow T_N$

従来法		推定クラス				再現率
		1.晴	2.曇	3.雨	4.雪	
正解クラス	1.晴	514	0	0	0	1.000
	2.曇	8	275	3	0	0.962
	3.雨	1	101	62	0	0.378
	4.雪	0	5	0	0	0.000
適合率		0.983	0.722	0.954	-	0.878

(従来法) SSL: $L_S + U_S + U_N \rightarrow T_S$

従来法		推定クラス				再現率
		1.晴	2.曇	3.雨	4.雪	
正解クラス	1.晴	624	0	0	0	1.000
	2.曇	25	331	87	0	0.747
	3.雨	2	46	168	0	0.777
	4.雪	0	1	2	6	0.667
適合率		0.959	0.876	0.653	1.000	0.874

(従来法) SSL: $L_S + U_S + U_N \rightarrow T_N$

従来法		推定クラス				再現率
		1.晴	2.曇	3.雨	4.雪	
正解クラス	1.晴	514	0	0	0	1.000
	2.曇	8	275	3	0	0.962
	3.雨	1	102	61	0	0.372
	4.雪	0	5	0	0	0.000
適合率		0.983	0.720	0.953	-	0.877

(提案法) CP-SSL: $L_S + U_S \rightarrow T_S$

従来法		推定クラス				再現率
		1.晴	2.曇	3.雨	4.雪	
正解クラス	1.晴	623	0	1	0	0.998
	2.曇	25	350	65	3	0.790
	3.雨	2	56	158	0	0.731
	4.雪	0	2	0	7	0.778
適合率		0.958	0.857	0.705	0.700	0.881

(提案法) CP-SSL: $L_S + U_S \rightarrow T_N$

従来法		推定クラス				再現率
		1.晴	2.曇	3.雨	4.雪	
正解クラス	1.晴	514	0	0	0	1.000
	2.曇	8	274	3	1	0.968
	3.雨	1	96	67	0	0.409
	4.雪	0	3	0	2	0.400
適合率		0.983	0.735	0.957	0.667	0.884

(提案法) CP-SSL: $L_S + U_S + U_N \rightarrow T_S$

従来法		推定クラス				再現率
		1.晴	2.曇	3.雨	4.雪	
	1.晴	623	0	1	0	0.998

正解クラス	2.曇	25	334	84	0	0.754
	3.雨	1	37	178	0	0.824
	4.雪	0	1	1	7	0.778
適合率		0.960	0.898	0.674	1.000	0.884

(提案法) CP-SSL: $L_S + U_S + U_N \rightarrow T_N$

従来法		推定クラス				再現率
		1.晴	2.曇	3.雨	4.雪	
正解クラス	1.晴	514	0	0	0	1.000
	2.曇	8	273	5	0	0.955
	3.雨	1	74	88	1	0.537
	4.雪	0	1	0	4	0.800
適合率		0.983	0.784	0.946	0.800	0.907

実験では、従来のラベルありデータのみを用いてモデルを構築するアプローチでは低頻度の「雪」をまったく推定することができなかった。従来の SSL については、ラベルありデータ (南関東) に南関東と北関東のラベルなしデータを組み合わせた場合に、南関東の「雪」を推定することできたが、それ以外の組み合わせでは推定することができなかった。

一方、提案手法では、いずれのデータセットの組み合わせでも「雪」を推定することができた。「雪」以外のクラスの推定精度については、データセットの組み合わせ方によって「雨」が従来法よりも精度が低くなった以外は、従来法と同程度かそれ以上の推定精度を達成した。ラベルありデータに組み合わせるラベルなしデータを「南関東のみ」と「南関東+北関東」とした場合、後者のほうがクラス分類精度が向上しており、SSL の効果が現れていることが分かる。

5. おわりに

不均衡マルチクラスデータに対する半教師あり学習 (semi-supervised learning (SSL)) について、ラベルありデータに対するクラス分類器による分割交差検証から得られる各ラベルありデータ点のクラス別分類確率を特徴量として用いる、SSL による疑似ラベル付与手法を提案した。提案手法は、従来の self-training とクラス分類器を組み合わせた手法と比べて、ラベルなしデータセットに対する疑似ラベル付与率、とくに低頻度クラスに対する疑似ラベル付与率を向上させる効果がある。ラベルあり・ラベルなし

本提案手法と SSL を用いない従来法と SSL の従来法の性能比較のために、関東 7 都市の各日の気象予報データを用いてその日の天気 (晴、曇、雨、雪の 4 クラス) を推定するクラス分類問題において、ラベルあり・ラベルなし・テスト用のデータセットの種々の組み合わせに提案手法と従来法を適用した。その結果、従来法では推定困難な、非常に低頻度の「雪」を提案手法で推定できることを確認した。また、「雪」以外のクラスの推定精度については、データセットの組み合わせ方によって「雨」が従来法よりも精度が低くなった以外は、従来法と同程度かそれ以上の推定精度を達成しており、本提案手法が不均衡マルチクラスデータセットの分類性能を向上するための手法として有望であることを示している。

本提案手法では、クラスが未知のデータ点は特徴量空間上で近くに分布するクラス分類確率のパターンが似たデータ点と同じクラスに属する可能性が高いという仮説を元にアルゴリズムを設計した。今後は、この仮説とアルゴリズムの妥当性と、今回の実験に用いた self-training と SVM 以外の組み合わせでの本提案手法の有効性についての検証に取り組む。

参考文献

[1] Alberto Fernandez, Salvador Garcıa, Francisco Herrera, Nitesh V. Chawla, “SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary”, Journal of Artificial Intelligence Research 61 (2018)

[2] Bassam A. Almogahed, Ioannis A. Kakadiaris, “Empowering Imbalanced Data in Supervised Learning: A Semi-supervised Learning Approach”, In: Wermter S. et al. (eds) Artificial Neural Networks and Machine Learning – ICANN 2014 (2014).

[3] 日本気象 (株), “気象データ配信> 提供データ一覧> 今日明日天気（市区町村）”, .https://n-kishou.com/corp/service/weather-data/data/catalog/tenki2.html

[4] Chih-Chung Chang, Chih-Jen Lin, “LIBSVM : a library for support vector machines”, ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, (2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm