不均衡・疎データ問題の対処法

内　田　　匠*・吉　田　健　一*

Approaches on imbalance and sparse data

Takumi UCHIDA* and Kenichi YOSHIDA*

1. はじめに

Web マーケティングデータは疎でありかつ不均衡になる傾向にあり、この二つの課題について同時に対処する必要がある。本研究では、この課題に対する一般的な対処法を先行研究を参照し、それぞれの対処法を組み合わせて適応した結果について比較した。不均衡データの対処法として、代表的な resampling, cost sensitive [1] を採用した。また、疎データの一般的な対処法として行列変換と変数選択を組み合わせて対処する。変数選択は wrapper method [2] を採用した。

2. 検証方法

始めに、疎かつ不均衡になるように疑似生成したデータと実際の Web マーケティングデータを用意する。次に、先行研究の対処法を組み合わせ各データに対して 10hold Cross Validation を行う。最後に、その組み合わせごとに出力された精度指標や処理時間を比較した。

3. 結　　果

3.1 不均衡データ対処法の単体の効果

疑似生成データでは、precision の値が悪化したが、recall が大きく改善し、f1-score も改善した。Web マーケティングデータでは、precision が悪化し recall は改善されたが、f1-score に大きな改善は見られなかった。

3.2 疎データ対処法の単体の効果

擬似生成データと Web マーケティングデータでも、各疎データ対処法で顕著な f1-score の改善は確認できなかった。また、SBS(Sequential Backward Selection) を適応した場合に f1-score が大きく悪化することが確認できた。疑似生成データでは、疎データに対して PCA で行列変換を施した場合に f1-score の改善が見られたが、Web マーケティングデータではその傾向は確認できなかった。

3.3 各対処法を組み合わせ際の結果について

SBS を単体で適応した場合に精度指標が悪化していたが、対処法を組み合わせた場合は精度が改善する傾向を確認できた。さらに、疑似生成データに決定木を適応する場合、不均衡データの対処法を適応しても recall の改善の度合いが比較的低いことが確認できた。しかし、under random resampling の場合は例外で、recall が改善され f1-score も他の手法と同程度に改善していた。これらの傾向は Web マーケティングデータにおいても確認出来た。

		feature_selection		
		平均 / f1-score		改善率 / f1-score
imbalance	convert_X	SBS	Nothing	SBS / Nothing
default	PCA	0.033	0.096	0.338
	SparseRandomProjection	0.000	0.059	0.000
	Nothing	0.000	0.071	0.000
over_resampling	PCA	0.194	0.266	0.729
	SparseRandomProjection	0.291	0.257	1.132
	Nothing	0.267	0.263	1.015
sample_weight	PCA	0.278	0.250	1.112
	SparseRandomProjection	0.294	0.238	1.235
	Nothing	0.283	0.287	0.985
smote	PCA	0.196	0.235	0.834
	SparseRandomProjection	0.280	0.253	1.107
	Nothing	0.254	0.269	0.943
under_resampling	PCA	0.265	0.215	1.237
	SparseRandomProjection	0.257	0.226	1.139
	Nothing	0.233	0.226	1.034

Fig. 1: 手法を組み合わせた時の SBS の F1 値改善度

3.4 疎かつ不均衡データに対処する時、留意すべきこと

不均衡データの一般的な対処法については、説明変数が疎であっても問題なく機能していた。一方で、疎データに対処法による明確な改善は確認できなかった。SBS は単体で適応しても 0 予測に偏り、精度悪化になることが確認できたが、不均衡データの対処法と組み合わせることで精度悪化を解消することができた。しかし、決定木は組み合わせる不均衡データの対処法でその精度に大きなばらつきがあった。

* 筑波大学大学院 ビジネス科学研究科
* Graduate School of Business Sciences, University of Tsukuba

参考文献

1) He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineer-ing, 21(9), 1263/1284.
2) Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16/28.