1S-04

モデルの適用範囲を考慮した半教師有り学習による不均衡データの分類

高宮 昂輝 †
関西学院大学 †

1 はじめに

現在の機械学習において，教師あり学習と呼ばれるものが主流であり，教師あり学習には正解ラベルが必要である．正解ラベルとは，あるタスクの入力データに対しての答えであり，例えばCT画像を用いて患者が病気であるか否かを分類するというタスクにおいて，各CT画像に対して，それを撮った患者が病気である・病気でないというラベルが正解ラベルとなる．正解ラベルを付ける作業には人手が必要であり，多くの数を用意することが難しい一方で，正解ラベルが付いていないデータ (ラベルなしデータ) は大量に用意できることが多い．そこで，正解ラベルが付いたラベルありデータに，ラベルなしデータを追加して利用することで，あるタスクの性能を上げるという学習方法が，半教師あり学習と呼ばれ，盛んに研究されている [2]．半教師あり学習はラベルありデータの数が少ない場合に特に有用であり，この状況に当てはまる例として，異常検知への利用が挙げられる．異常検知では教師データとして与えられる異常データの数が限られている場合が多く，正常データについて教師データの数が少ないことがあるため，半教師あり学習により利用できるデータを増やすことは有意義である．

本研究では, 半教師あり学習の中でも，ラベルありデータから作成した分類器を利用して，ラベルなしデータに仮のラベルを付けて利用するアルゴリズムである，自己訓練法というアルゴリズムの拡張および異常検知への利用を目指す．提案手法では，ラベルなしデータを全て利用するのではなく，利用するか否かを判定する指標を作成し，その指標に合うデータのみを利用して半教師あり学習を行う．

Classification of imbalanced data by semi-supervised learning considering the applicable range of the model
† Koki Takamiya, Kwansei Gakuin University

2 自己訓練法

自己訓練法は，半教師あり学習の中でも最も基本的な手法である．この手法では，最初にラベルありデータから分類器を作成し，その分類器を用いてラベルなしデータを分類した結果を仮の正解ラベルとして付与する．その後，ラベルありデータと仮の正解ラベルが付与されたラベルなしデータにより分類器を再学習して，再度ラベルなしデータの分類を行い，仮のラベルと分類器を更新する，という学習を一定回数，もしくは収束するまで繰り返す．自己訓練法は，1995年にYarowsky によって，テキスト文書の分類のためのアプローチとして初めて提案された [3]．それ以来，様々な応用がなされている [1]．

3 提案手法

通常の自己訓練法では，ラベルなしデータを誤分類してしまい，誤ったラベルをつけてしまうことで，分類性能が下がってしまうことが問題となっている．そこで，データの生成モデルとして，混合ガウス分布モデルを考え，ラベルなしデータの生成確率を踏まえた上で，仮のラベルを付与して利用するかどうかを検討する，自己訓練法の拡張を目指す．提案手法では，全てのラベルなしデータを利用するわけではなく，ラベルなしデータのうち生成確率がしきい値以上かつ分類確率が一定範囲のものを抽出し，それらにのみ仮のラベルを付与して利用する．典型的な自己訓練法の拡張では，分類器として分類モデルを利用し，ラベルなしデータの利用の判定を，生成確率でなく推定された密度比 (分類確率) のみによって行っている物が多かった．この場合，密度比は大きくとも生成確率は低い，ということが起こり，

本来信用するべきでない不確実性の高い情報に対しても，過剰な自信を抱いてしまい，仮のラベルを付与して利用する，という間違いが発生する一方，提案手法では，生成モデルを利用することにより，生成確率を利用することが出来るようになり，過剰な自信を抱くのを防止している．また，正常データと異常データの数に偏りがあるという異常検知の問題点を解決するため，正常データについては分類確率が低いものを除いている．

4　実験と考察

半教師あり学習の有効性を確かめるため，wilt dataset および pima dataset を用いて実験を行った．wilt dataset において，ラベルありデータとして，異常データを 10 個，正常データを 100 個，ラベルなしデータとして，異常データを 30 個，正常データを 300 個，pima dataset において，ラベルありデータとして，異常データを 10 個，正常データを 20 個，ラベルなしデータとして，異常データを 30 個，正常データを 60 個用いて，混合数 2 の混合ガウス分布による学習を行った．利用するデータかどうかを判定する生成確率のしきい値は，10^{-6}，10^{-13}，分類確率のしきい値は 0.8，0.99 をそれぞれ利用し，学習回数は 20 回とした．評価指標としては，ROC 曲線の下部面積である AUC を使用した．

表 1 に実験の結果を示す．性能を見ると，ラベルなしデータを利用しない場合，つまり単純な教師あり学習の性能が最も低くなっており，半教師あり学習を下回った．このことから，教師データが少ない場合，半教師あり学習自体に性能を向上させる効果があることが確認できる．また，提案手法の性能は，教師あり学習や他の半教師あり学習よりも高い性能を示している．提案手法では，一定の基準を用いて利用するデータを制限し，誤分類してしまうようなラベルなしデータを利用しないようにするため，通常の半教師あり学習や教師あり学習に比べて，性能の向上が見込まれることがわかる．図 1 に教師あり学習と提案手法 (半教師あり学習) の pima dataset に対する ROC 曲線を示す．この図を見ると，偽陽性率 (横軸) が 0.2 程度の時点から性能の向上が確認できるため，偽陽性率を高くすることができないようなケースにおいても有効であると考えられる．

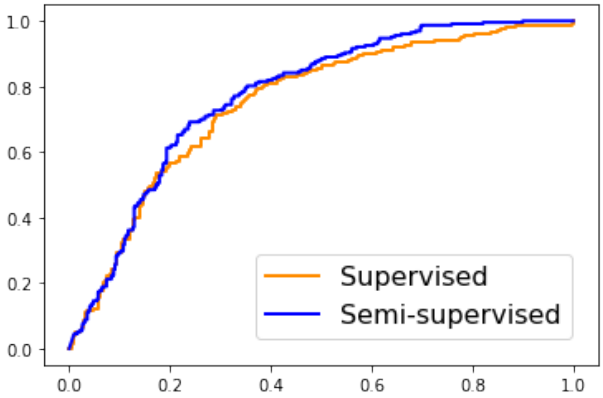

図 1: 教師あり学習と半教師あり学習の比較

手法	wilt	pima
教師あり	0.806	0.750
半教師あり (全てのデータを利用)	0.825	0.737
半教師あり (分類確率を利用)	0.822	0.766
半教師あり (生成確率を利用)	0.829	0.767
提案手法	0.832	0.775

表 1: 実験結果

5　まとめ

本稿では，異常検知問題において，半教師あり学習を適用し精度の向上を目指し，自己訓練法を拡張した，生成確率と分類確率という二つの基準を突破したデータのみを使用するモデルを提案し，評価を行った．提案手法は異常データの数が 10 個程度と少ない場合に，他の手法を AUC において上回った．

参考文献

[1] Triguero, I., García, S. and Herrera, F.: Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study, *Knowledge and Information systems*, Vol. 42, No. 2, pp. 245–284 (2015).

[2] Van Engelen, J. E. and Hoos, H. H.: A survey on semi-supervised learning, *Machine Learning*, Vol. 109, No. 2, pp. 373–440 (2020).

[3] Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods, *33rd annual meeting of the association for computational linguistics*, pp. 189–196 (1995).