05 低い誤検知率で異常を検知

部分AUC最大化のための半教師あり学習

どんな研究

機械学習における二値分類問題において、ラベルなしデータを活用することによって、部分AUC（偽陽性率が特定の範囲での真陽性率）を高める分類器の学習方法に関する研究です。例えば異常検知において、誤検知率を低く抑えた状態で真の異常を見逃さなくすることが可能になります。

どこが凄い

従来技術ではラベルなしデータを活用することができませんでした。私たちはラベルありデータとラベルなしデータを用いて近似的な部分AUCを計算する方法を考案しました。そして、その近似部分AUCを最大化させるように分類器を学習することによって、高精度の分類を可能にしました。

めざす未来

本技術を用いることで、コストが高いラベルを付ける作業を減らしても高い分類性能を達成することができます。これまで少数の学習データしか得られない場合には機械学習技術を適用できませんでしたが、本技術を発展させることにより、機械学習の適用範囲を広げることをめざします。

部分AUC

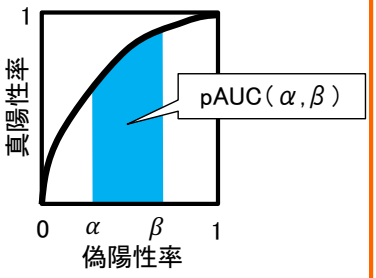
PAUC: Partial Area Under the Receiver Operating Characteristic Curve

PAUC： $\alpha \leq$ 偽陽性率 $\leq \beta$ のときの真陽性率曲線（ROC）の下の面積

正例の割合と負例の割合が大きく異なる二値分類問題の評価尺度として注目

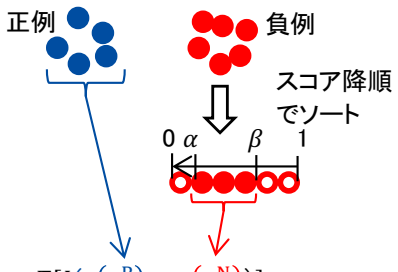
応用例1：異常検知：異常と誤って検知するとオペレータの負担増

応用例2：病気診断：病気と誤って診断すると不必要な検査によりコスト増



従来技術（教師あり学習）

ラベルありデータ（正例、負例）を入力とし、部分AUC最大化により正例らしさを出力するスコア関数sを学習



$$\mathbb{E}[I(s(\mathbf{x}^{\mathrm{P}}) > s(\mathbf{x}^{\mathrm{N}}))] = \mathrm{pAUC1}(\alpha, \beta)$$

正例と負例で計算される部分AUC

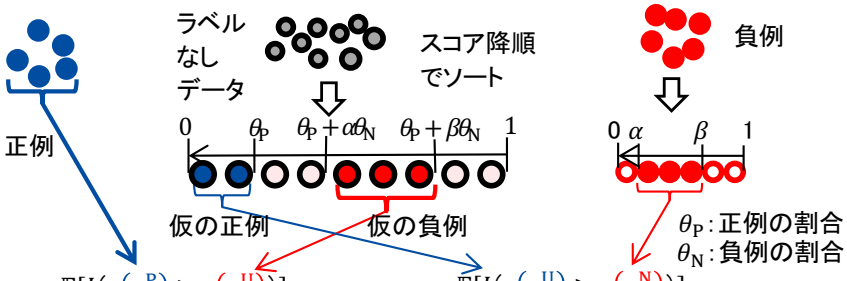
目的関数

$$L = \mathrm{pAUC1}(\alpha, \beta)$$

提案技術（半教師あり学習）

ラベルなしデータから計算できる2つの部分AUCの近似を導出し、スコア関数の学習のために利用

正例の割合θ_{p}を境に、スコアの良い（悪い）ラベルなしデータは仮の正例（負例）とみなす



$$\mathbb{E}[I(s(\mathbf{x}^{\mathrm{P}}) > s(\mathbf{x}^{\mathrm{U}}))] \approx \mathrm{pAUC2}(\alpha, \beta)$$

正例とラベルなしデータで計算される近似部分AUC

$$\mathbb{E}[I(s(\mathbf{x}^{\mathrm{U}}) > s(\mathbf{x}^{\mathrm{N}}))] \approx \mathrm{pAUC3}(\alpha, \beta)$$

負例とラベルなしデータで計算できる近似部分AUC

目的関数：3つの部分AUCの重み付き和

$$L = \lambda_1 \mathrm{pAUC1}(\alpha, \beta) + \lambda_2 \mathrm{pAUC2}(\alpha, \beta) + \lambda_3 \mathrm{pAUC3}(\alpha, \beta)$$

関連文献

[1] T. Iwata, A. Fujino, N. Ueda, "Semi-supervised learning for maximizing the partial AUC," in *Proc AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

連 絡 先

岩田 具治（Tomoharu Iwata） 上田特別研究室
Email: cs-openhouse-ml@hco.ntt.co.jp