共同学習における分類器の合意度を用いた追加データ選択法の提案

An unlabaled data selection method based on the agreement among classifiers in Co-Learning

岡谷 一宏　　吉田 哲也
Kazuhiro Okatani　Tetsuya Yoshida

北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

We propose an unlabaled data selection method based on the agreement among classifiers in Co-Learning. In semi-supervised learning, the selection of unlabaled data, which are to be used as pseudo-labeld data, affects the classification accuracy of the constructed classifier. When unlabeled data with disagreed predicted labels among classifieres are selected and utilized, the size of the selected unlabeld data can become too large and degrade the performance. Our method partitions the unlabeled dataset based on the number of agreed classifiers, and incrementally adds the subset into the candidate pseudo-labeled dataset in descending order of the agreement. We evaluated the proposed method against UCI repository, and report the results.

1. はじめに

教師あり学習を行う際，求めたい情報（クラスラベル）の付随しているデータ（ラベルありデータ，$\boldsymbol{L}$ と表記）をあらかじめ用意する必要がある．基本的に $\boldsymbol{L}$ が多いほど高精度の分類が可能になるが，データ一つ一つにクラスラベルを与えて用意するにはコストがかかる．この問題に対し，クラスラベルの付随していないデータ（ラベルなしデータ，$\boldsymbol{U}$ と表記）も学習時に使う半教師あり学習が提案された [1]．ラベル付けの手間を軽減でき，$\boldsymbol{U}$ も含めた大量のデータを活用することで，$\boldsymbol{L}$ のみの場合よりも高精度な分類が可能になると期待される．

本稿では，半教師あり学習の一つである共同学習において分類器間の合意度に基づいて $\boldsymbol{U}$ の部分集合を選択する手法を提案する．合意度を用いることで，疑似ラベルを与えた $\boldsymbol{U}$ から，より信頼性の高い疑似ラベルが与えられたデータを優先して使用することが可能になると期待される．

2. 共同学習

共同学習とは半教師あり学習の一つで，複数の分類器を用いて互いに学習しあう学習法である．

共同学習では，まず $\boldsymbol{L}$ を用いて複数の分類器を作成し，$\boldsymbol{U}$ をそれぞれの分類器に入力してクラスラベルを出力させる．出力されたクラスラベルを基に $\boldsymbol{U}$ を部分集合に分割し，部分集合から分類器構築に使用する追加データ $\boldsymbol{L}_j$（分類器 h_j への追加データ）を選択する．そして，$\boldsymbol{L} \cup \boldsymbol{L}_j$ を用いて各分類器を再構築する．再構築した分類機で $\boldsymbol{U}$ にクラスラベルを与え部分集合を選択し，分類器の再構築する，という手順を，全ての $\boldsymbol{L}_j$ が更新されなくまで繰り返す (Algorithm1 参照)．

これまでに様々な追加データ選択法が提案されてきた．文献 [2] では，$\boldsymbol{U}$ を分類器で分類し，決定木の同じ葉の部分に分類されたデータ（同値類）ごとに分割し，そのうちの一つの部分集合を選択する．しかし，使用する学習アルゴリズムが決定木を作るものに限定されるという課題がある．文献 [3] では，３つの分類器の出力において，データを追加される分類器以外の２つの分類器の出力が同じになるデータ集合を選択する．選択したデータ集合が大きすぎるとランダムにサンプリングして数を減らしたデータ集合を選択データ集合とする．しかし，選択されるデータ集合が大きすぎてランダムにサンプリングする場合が多くなり，適切なデータ集合を選択できない恐れがある．

連絡先: 岡谷 一宏，北海道大学大学院情報科学研究科 コンピュータサイエンス専攻, TEL:011-706-7260, E-mail okatani@meme.hokudai.ac.jp

Algorithm 1 共同学習アルゴリズム

Require: $\boldsymbol{U}$ //ラベルなしデータ集合
$\boldsymbol{L}$ //ラベルありデータ集合
$A_i, ..., A_r$ //学習アルゴリズム

1: **for** each learning algorithm A_i **do**
2: $\boldsymbol{L}_i \leftarrow \emptyset$ //A_i への追加データ
3: $h_i \leftarrow A_i(\boldsymbol{L})$
4: **end for**
5: **while** $\boldsymbol{L}_1, ..., \boldsymbol{L}_r$ change **do**
6: $\boldsymbol{C} \leftarrow$ DivideData($\boldsymbol{U}$,$\boldsymbol{h}$)
7: $\{ \boldsymbol{L}_1, .., \boldsymbol{L}_r\} \leftarrow$SelectData($\boldsymbol{L}$,$\boldsymbol{C}$)
8: $\boldsymbol{h} \leftarrow$ RebuildClassifiers($\boldsymbol{L}$,$\{ \boldsymbol{L}_1, .., \boldsymbol{L}_r\}$,$\boldsymbol{h}$)
9: **end while**
10: $h \leftarrow$ CombineClassifiers($\boldsymbol{h}$) // 本稿では省略

3. 合意度を用いた追加データ選択法

本稿では，$\boldsymbol{U}$ からの部分集合の選択法として合意度を用いた追加データ選択法を提案する．各データに対し，同じクラスラベルを出力する分類器の数を合意度と定義し，合意度に基づいて $\boldsymbol{U}$ を部分集合に分割する（図 1 参照）．この理由は，分類器の合意度が高いデータほどより多くの分類器が同一のクラスラベルを出力しているという意味で予測クラスラベルの信頼性が高いと考え，より信頼性の高い疑似クラスラベルを付加したデータを優先的に使用するためである．部分集合間の包含関係に基づいて合意度の高いものから逐次的に拡張することにより，疑似クラスラベルの信頼性が高く，学習に効果的なデータ集合を選択できると期待される．

提案手法では，まず分類器を r（≥ 3）個用意し，従来手法と同様に $\boldsymbol{U}$ に各分類器でクラスラベルを与える．次に，合意度ごとに部分集合に分割する（Algorithm2 参照）．そして，合意度の降順に追加データとして選択し，精度向上の期待できる

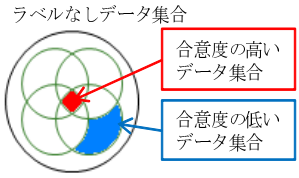

図 1: 合意度に基づくデータ分割

合意度のデータのみを追加データとして使用し（Algorithm3 参照），分類器を再構築する（Algorithm4 参照）．

Algorithm 2 DivideDataByAgreement($\boldsymbol{U}$,$\boldsymbol{h}$)

Require: $\boldsymbol{U}$ //ラベルなしデータ集合
$h_1, ..., h_r \in \boldsymbol{h}$ //分類器集合
/*————-U を部分集合に分割————-*/
for each unlabeled example $x \in \boldsymbol{U}$ **do**
  for possible labels k **do**
    $c_k = |\{h_j | h_j(x) = k\}|$ //k を出力する分類器数
  end for
  $c = \arg\ \max_k\{c_k\}$ //最も多く出力されるクラス値
  for each classifier h_i **do**
    if $h_i(x) \neq c$ **then**
      $a = \max\{c_k\}$ //合意度 (c を出力する分類器数)
      $\boldsymbol{L}_i^a \leftarrow \boldsymbol{L}_i^a \cup \{(x, c)\}$
    end if
  end for
end for

Algorithm 3 SelectDataByAgreement($\boldsymbol{L}$,$\boldsymbol{L_i^a}$)

Require: $\boldsymbol{L}$ //ラベルありデータ集合
$\boldsymbol{L}_i^a$ //分割データ集合
/*————-合意度の降順に追加データ選択————-*/
for each classifier h_i **do**
  for $\boldsymbol{L}_i^a$ in descending order of a **do**
    $e_i^a \leftarrow MeasureError(h_k)$//(k ≠ i)
    $\boldsymbol{L}_i \leftarrow \boldsymbol{L}_i \cup \boldsymbol{L}_i^a$
    $q_i = |\boldsymbol{L} \cup \boldsymbol{L}_i|(1 - \frac{2e_i^a}{|\boldsymbol{L} \cup \boldsymbol{L}_i|})^2$ //データ追加時の評価値
    if $q_i > q_i'$ //q_i' : データ追加前の評価値 **then**
      $update_i \leftarrow TRUE$
      break
    end if
  end for
end for

4. 評価実験

4.1 実験設定

提案手法を UCI repository のデータセットに対して評価した．学習アルゴリズムとして，Naive Bayes，k-nearest neighbor，J48, K-Star，Random Forest，Logistic，Multilayer Perceptron の計 7 個用いた．データセット $\boldsymbol{D}$ に対し，$\frac{1}{4}|\boldsymbol{D}|$ をテストデータ ($\boldsymbol{T}$ と表記) とし，残りのデータの $\boldsymbol{L}$ と $\boldsymbol{U}$ の割合を変えて実験を行った．

Algorithm 4 RebuildClassifiers($\boldsymbol{L}$,$\boldsymbol{L_i}$,$update_i$)

Require: $\boldsymbol{L}$ //ラベルありデータ集合
$\boldsymbol{L_i}$ //追加データ集合
/*————-合意度の降順に追加データ選択————-*/
for each classifier h_i **do**
  if $update_i = TRUE$ **then**
    $q_i' \leftarrow q$
    $h_i \leftarrow A_i(\boldsymbol{L} \cup \boldsymbol{L}_i)$
  end if
end for

4.2 実験結果

Data set	initial	final	improv (%)	exs from U labeled	avg. #rounds
colic	.8119	.8195	4.21	13.3	5.3
diabetes	.7390	.7442	2.03	40.0	7.9
ionosphere	.8655	.8770	9.35	10.5	4.7
tic-tac-toe	.8719	.8870	13.36	53.8	8.1
vote	.9453	.9500	9.40	5.4	4.1

表 1: エラー率（10 回平均）

$\boldsymbol{L}$: $\boldsymbol{U}$=1:4 の場合の結果を表 1 に示す．表 1 より，全てのデータにおいて分類精度が向上することを確認した．しかし，$\boldsymbol{L}$ と $\boldsymbol{U}$ の割合によっては精度が低下する場合もあった．特に，$\boldsymbol{L}$ の割合が少なすぎる場合に精度低下の傾向が見られた．これは，現状では $\boldsymbol{L}$ に基づいて追加データ選択時のノイズ率の推定を行っているが，推定に使用するデータ量が十分でないために推定値の信頼性が低く，真のクラスラベルとは異なる疑似クラスラベルが付与されたデータを追加しているためと考えられる．

5. まとめ

本稿では，共同学習における分類器の合意度を用いた追加データ選択法を提案した．合意度を用いて $\boldsymbol{U}$ をより細かな部分集合に分割することにより，疑似クラスラベルの信頼性が高いと期待されるデータ集合を選択する手法を提案した．しかし，選択した追加データ集合におけるノイズ率の推定が不十分であるという課題があるため，今後はこの課題に対し取り組んでいく予定である．

参考文献

[1] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with to-training. In *Proc.11th annual Conf. on Computational Learning Theory*, pp. 92–100, 1998.

[2] Sally Goldman and Yan Zhou. Enhancing supervised learning with unlabeled data. In *Proc.17th Int. Conf. on Machine Learning*, pp. 327–334, 2000.

[3] Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. In *IEEE Trans. on Knowledge and Data Engineering, vol.17*, pp. 1529–1541, 2005.