1T-7

能動学習による Tri-Training 手法の改善*

小野 隆太†
東北大学工学部情報工学科

寺邊正大‡
東北大学大学院情報科学研究科

橋本和夫§
東北大学大学院情報科学研究科

1 はじめに

Web ページをその内容に基づき自動分類する様々な手法が提案されている．Blum らは，半教師あり学習の代表的な手法の 1 つである Co-Training 手法 [1] を提案した．一般に，学習に用いられるラベルあり事例が多いほど，導出される分類モデルの分類精度は向上する．Co-Training 手法は 2 つの分類モデルを用いて，ラベルなし事例にラベル付けを行い，ラベルあり事例を増やすことで，分類モデルの分類精度を高めるものである．Co-Training 手法は Web ページ分類において，通常の教師あり学習よりも高い分類精度を持つ分類モデルを導出すると報告されている．しかし，Co-Training 手法は分類モデルが付与する確信度の高い順にラベル付けを行うため，本来ラベル付けを行いたい，分類モデルがラベルの分類を間違いやすいラベルなし事例には十分にラベル付けを行うことができない．

Muslea らは，Co-Training 手法よりもラベル付けにコストを費やす代わりに，Co-Training 手法よりも高い分類精度を持つ分類モデルを導出する Co-Testing 手法 [2] を提案した．Co-Testing 手法は，Co-Training 手法に能動学習を導入した手法である．能動学習とは，指定した範囲から学習に用いるデータをサンプリングする手法である．Co-Testing 手法は，分類モデルがラベルの分類を間違いやすいラベルなし事例のみを選択し，これに対して人間が正しいラベルを教示するため，少数のラベルなし事例により分類精度の大きな改善を可能にした．

Zhou らは，3 つの分類モデルによりラベルなし事例にラベル付けを行い，それらをラベルあり事例集合と共に学習に用いて 3 つの分類モデルを再導出する Tri-Training 手法 [3] を提案した．Tri-Training 手法は学習フェイズが終了すると，分類フェイズにて 3 つの分類モデルの多数決により新たな事例を分類する．

Tri-Training 手法は，3 つの分類モデルのうち，少なくとも 2 つの予測が一致したラベルなし事例へラベル付けを行うが，実際は全ての分類モデルの予測が一致している事例にラベル付けを行っていることが多い．そのため，Tri-Training 手法は Co-Training 手法同様，分類フェイズにて分類を誤りやすいラベルなし事例へのラベル付けを十分に行うことができない．

本研究では，Tri-Training 手法に能動学習を導入した Tri-Testing 手法を提案する．能動学習を用いることにより，分類精度の改善に大きく寄与する少数のラベルなし事例を人間が教示することで，Tri-Training 手法よりも高い分類精度を持つ分類モデルの導出が可能となる．

2 学習に有用な事例のサンプリング

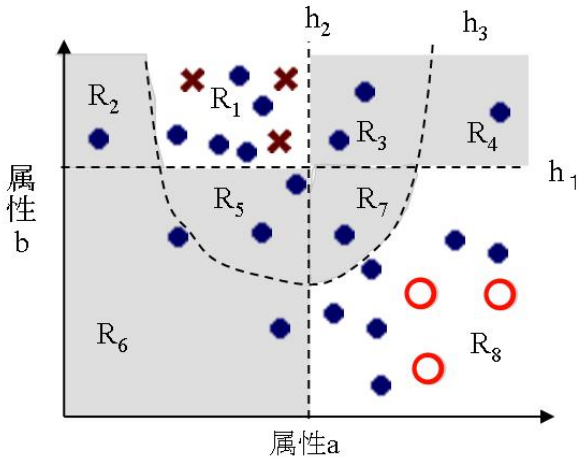

図 1: 3 つの分類モデルによる領域の分割

Tri-Training 手法において分類精度の改善に有用なラベルなし事例の説明を行う．まず，説明を簡単にするため，属性が a と b の 2 つしかなく，ラベルも○と×の 2 種類であるデータについて考える．横軸を属性 a の値，縦軸を属性 b の値として，ラベルあり 6 事例とラベルなし 20 事例をプロットした結果，図 1 のようになったとする．さらに，学習を行った結果，3 つの分類モデル h_1，h_2，h_3 が図 1 のように導出されたとする．この 3 つの分類モデルによって，図 1 は R_1 から R_8 の 8 つの領域に分けられる．この中で，R_1 と R_8 に含まれる事例は現時点の全ての分類モデルのラベルの予測が一致している事例である．そのため，R_1 と R_8 に含まれる事例は分類フェイズにて，ほとんどの事例が正しく分類される．従って，R_1 と R_8 に含まれるラベルなし事例に新たにラベル付けをして，この事例を加えて再度学習を行っても分類精度はあまり向上しない．よって，R_1 と R_8 に含まれる事例は，コストをかけてラベル付けをするには値しない．

一方，R_2 から R_7 に含まれる事例は，3 つの分類モデルのうちどれか 1 つの分類モデルの予測が異なっている事例

*Improvement of Tri-Training using active learning
†Ryuta Ono, Department of Information Engineering, School of Engineering, Tohoku University
‡Masahiro Terabe, Graduate School of Information Science Engineering, Tohoku University
§Kazuo Hashimoto, Graduate School of Information Science Engineering, Tohoku University

である．現時点では R_2 から R_7 に含まれる事例は分類フェイズで誤って分類されることが多い．しかし，このような事例に正しいラベルを付与することができ，さらにこれを学習に用いることができれば，分類精度を少数の事例で大きく改善することが可能である．すなわち，R_2 から R_7 に含まれるラベルなし事例は，分類精度の改善に有用なラベルなし事例であるということができる．また，どちらの場合でも，誤ったラベルが付与された事例を学習に用いると，分類精度が低下してしまうことになる．

以上より，少数の事例で大きく分類精度を改善するには，R_2 から R_7 に含まれる事例に，正しいラベルを付与した事例を学習に用いれば良いと言える．これは用いるデータの属性数が増加しても同様である．

3 Tri-Training 手法の改善

Tri-Training 手法のラベル付け方法は，2 つの分類モデルの予測が一致した事例にラベル付けをするというものである．しかし実際は，全分類モデルの予測が一致している事例，すなわち R_1 と R_8 に含まれる事例にラベル付けを行っていることが多い．このようなラベル付け方法には，分類精度をあまり改善しないラベルなし事例にばかりラベル付けをしていることに加え，誤ってラベル付けをしてしまう可能性もある．

そこで，R_2 から R_7 に含まれる任意の数の事例を選択し，人間による正しいラベル付けを行う手法として，Tri-Testing 手法を提案する．Tri-Testing 手法は，人間がラベルなし事例のラベルを教授するため，Tri-Training 手法よりもコストがかかる．ただし，Tri-Testing 手法はラベル付けをするラベルなし事例を分類精度の改善に大きく寄与する事例に限定するため，ラベル付けにかかるコストを最小化しつつ，高い分類精度を持つ分類モデルを導出する．このように，Tri-Testing 手法は Tri-Training 手法よりもラベル付けにコストを要する代わりに，Tri-Training 手法よりも高い分類精度を持つ分類モデルを導出するための手法である．

Tri-Testing 手法のアルゴリズムの流れは表 1 に記述した．L をラベルあり事例集合，U をラベルなし事例集合，$Learn$ を学習アルゴリズム，h_i を分類モデル，S_i を事例数はそのままに L の事例構成を変えたラベルあり事例集合，$ContentionPoints[i]$ を R_2 から R_7 に含まれる事例を記憶するリスト，L_i をラベル付けされたラベルなし事例集合とする．また，SelectQuery() は h_i による確信度が最も高い事例 x_i を返すという関数である．Tri-Testing 手法は，ラベル付けをした後にラベル付け結果 $h_i(x_i)$ と人間が付与した正しいラベル y_i の比較を行う．誤った予測をされていた事例に正しいラベルを教授することにより，その事例を誤って予測していた分類モデルにラベル付けした事例を学習させることが可能となる．そのため，Tri-Testing 手法は Tri-Training 手法よりも高い分類精度を持つ分類モデルを導出可能であると考えられる．

表 1: Tri-Testing 手法の擬似コード

Input: L, U, $Learn$
for $i \in \{1..3\}$ **do**
　$S_i \leftarrow BootstrapSample(L)$ //復元抽出法
　$h_i \leftarrow Learn(S_i)$ //h_i の導出
end of for
for k iterations //k は任意数
　for $i \in \{1..3\}$ **do**
　　for every $x \in U$ **do** $(j, k \neq i)$
　　if $h_j(x) = h_k(x)$ and $h_j(x) \neq h_i(x)$
　　then $ContentionPoints[i] \leftarrow x$
　　end of for
　　$x_i \leftarrow SelectQuery(ContentionPoints[i])$
　　$y_i \leftarrow Label(x_i)$ //人間によるラベル付け
　　if $y_i \neq h_i(x_i)$
　　then $L_i \leftarrow L_i \cup \{(x_i, y_i)\}$
　　else then $L_j, L_k \leftarrow L_j, L_k \cup \{(x_i, y_i)\}$
　end of for
　for $i \in \{1..3\}$ **do**
　　$h_i \leftarrow Learn(L \cup L_i)$
　end of for
end of for k iterations
Output: $h(x) \leftarrow \arg \max_{y \in label} \sum_{i:h_i(x)=y} 1$ //多数決

4 まとめ

Tri-Training 手法は，学習に有用なラベルなし事例に正しいラベル付けをすることが十分にできなかった．本論文では，学習に有用な事例のみを能動的に選択し，人間がラベルを教示し，これを学習に利用する Tri-Testing 手法を提案した．Tri-Testing 手法は，Tri-Training 手法よりもラベル付けにコストはかかるものの，教示コストを最小限に抑えつつ，Tri-Training 手法よりも高い分類精度を持つ分類モデルを導出する．今後，Tri-Testing 手法と Tri-Training 手法の比較実験を行い，各手法による分類モデルの分類精度とラベル付けにかかる教示コストを比較する予定である．

参考文献

[1] A. Blum, T. Mitchell, “ Combining labeled and unlabeled data with co-training ”, In *Proceedings of the 11th Annual Conference on Computational Learning Theory, Madison, WI*, pp.92-100, 1998.

[2] I. Muslea, S. Minton, C. A. Knoblock, “ Selective Sampling With Naive Co-Testing: Preliminary Result ”, In *The ECAI-2000 workshop on Machine Learning for information extraction*, 2000.

[3] Z. H. Zhou, M. Li, “ Tri-training: exploiting unlabeled data using three classifiers ”, In *IEEE Trans.Knowledge and Data Engineering*, vol.17, pp.1529-1541, 2005.