Hoeffding Tree Based Dynamic Integration of Classifiers for Ensemble Learning

西村 聖† 寺邊 正大‡ 橋本 和夫†

1 はじめに

近年，インターネットの普及やセンサ技術の発達にともない，データストリームからの分類学習が注目されている．データストリームでは時間経過にともない，コンセプトドリフトと呼ばれる学習対象 (コンセプト) や，対象から得られるデータ傾向の変化が発生する．そのため，データストリームからの分類学習では，分類器を最新のコンセプトへと適宜追従させる必要がある．

データストリームからの分類学習では，分類器群によるアンサンブルを構成する方法 [1, 2, 3] が一般的である．データストリームを一定サイズのチャンクに分け，チャンクごとに分類器を学習する．そして，過去に学習した分類器群のうち，最新のコンセプトと一致する分類器を用いてアンサンブルを構成する．

分類器の最新のコンセプトへの一致性を図る指標としては，最新のチャンクに対する精度が用いられる．この際，コンセプトドリフトの影響は属性空間ごとに異なるため，分類器の局所的な精度を求め，属性空間ごとに異なるアンサンブルを構成することで精度を高めることができる [2, 3].

Tysmbal らは，k 個の近傍事例を用いて分類器の局所的な精度を求めている [3]．しかし，彼らの手法はノイズが大きい場合や事例数が少ない場合は，正確に分類器の局所的な精度を求めることが困難となる．

一方，Zhu らは分類器の予測結果に関するメタ知識を学習し，分類器の局所的な精度を求めている [2]．分類器の予測結果に関するメタ知識を学習することで，ノイズや事例数不足による影響を抑制できる．しかし，彼らの手法は各属性の独立性を仮定しているという問題がある．

本論文では，データストリームからの決定木学習法である Hoeffding Tree[4, 5] を用いた，新たなアンサンブル学習法を提案する．Hoeffding Tree はデータストリームから高精度な決定木を学習可能である．そこで，分類器の予測結果に関するメタ知識を Hoeffding Tree を用いて学習することで，正確に分類器の最新のコンセプトへの一致性を判断できる．その結果，アンサンブル全体の精度が改善される．

†東北大学大学院 情報科学研究科，Graduate School of Information Sciencies, Tohoku University
‡株式会社 三菱総合研究所，Mitsubishi Resarch Institute, Inc.

2 提案手法

提案手法の学習アルゴリズムを表1に示す．提案手法は K 事例が到着するごとに，新しく分類器 C_i と，C_i の予測結果に関するメタ学習器 M_i を学習する．ここで，分類器やメタ学習器の総数が多くなると処理時間が大きくなるため，直近 N 個のみを保持することとする．なお，本論文ではメタ学習器として UFFT[5] を採用する．

表 1: The Learning Algorithm of Proposal Method.

Input: S : Data Stream.
K : Chunk Size.
Output: C : Recent N Classifiers.
M : Meta-Learner for each classifier.

Procedure

$C = \Phi$, $M = \Phi$, $\Delta = \Phi$, $i = 1$
For each arrival of new example $\vec{e}$ from S
  $\Delta = \Delta + \vec{e}$
  if $|\Delta| > K$
    $C_i = LearnClassifier(\Delta)$
    $C = C + C_i$
    if $|C| > N, C = C - C_{i-N-1}$
    $M_i = CreateNewMetaLearner()$
    $M = M + M_i$
    if $|M| > N, M = M - M_{i-N-1}$
    $\Delta = \Phi$, $i++$
  For each Meta-Learner $M_j \in M$
    $UpdateMetaLearner(M_j, C_j, \vec{e})$

また，新規事例 $\vec{e}$ の到着時には，新規事例を分類器 $C_j \in C$ で分類し，メタ学習器 $M_j \in M$ の更新を行う．多くの事例を学習に用いるほど，メタ学習器は正確に分類器の予測結果について学習できる．しかし，コンセプトドリフト以前の古い事例を学習に用いると，メタ知識を正確に学習することが困難となる．一方，Hoeffding Tree はコンセプトドリフトの有無を考慮しながら逐次的に決定木を学習できる．そのため，メタ学習器として Hoeffding Tree を用いることで，分類器の予測結果に関するメタ知識を正確に学習できる．

テスト事例の分類時には，テスト事例をメタ学習器に適用し，テスト事例に関して正しく学習できている分類器を求める．そして，それらの分類器によるアンサンブ

ルを構成し，テスト事例の分類を行う．

3 実験

3.1 実験データ

提案手法の性能を評価するため，Hulten らの人工データの作成方法 [4] を参考に，人工的にコンセプトドリフトを含むデータストリームを生成した．各事例 $\vec{e}$ のクラスは式 (1) で表わされる属性空間上の超平面にしたがい決定される．

$$\sum_{i=1}^{d} w_i e_i = w_0. \quad (1)$$

ここで w_i は重みベクトル $\vec{w}$ の i 番目の値であり，i 番目の属性のクラスの判定への寄与度を示している．コンセプトドリフトは定期的に重みベクトルを変化させ，超平面を変化させることにより実現される．詳しい説明は [4] を参照されたい．

3.2 実験方法

上記の方法により，人工的に学習事例 25 万からなるデータストリームを作成し，学習事例 2500 毎にテスト事例 1000 を与え分類精度を調べた．チャンクのサイズは 5000 とし，学習事例 5000 毎に C4.5 を用いて新たに分類器を学習する．保持する分類器の総数 N は 20 とし，直近 20 個の分類器のみを保持する．

また，コンセプトドリフトは学習事例 5 万毎に発生させており，コンセプトドリフトのたびにおよそ 10%の事例のクラスが変化している．学習事例 2500 毎にテストを行うため，合計で 100 回のテストが行われ，以下に示す結果はこの 100 回のテストでの平均値である．

3.3 比較手法

比較手法には，Tsymbal らの手法 (NN-Ensemble)[3] と，Zhu らの手法 (AO-DCS)[2] を用いた．本来は，NN-Ensemble は精度が高い $N/2$ 個の分類器を分類に用い，AO-DCS は最も精度が高い分類器のみを分類に用いる．しかし，本論文では Zhang らの手法 [1] を用いて，動的にアンサンブルに取り入れる分類器数を調整することとする．なお，各手法のパラメータはデフォルトを用いた．

3.4 実験結果

各手法の分類精度の属性数変化を図 1 に示す．チャンクサイズは 5000 と固定してあるため，属性数が大きくなるほど事例は属性空間上に疎に分布し，学習が難しくなる．

今回作成したデータは各属性の独立性が成り立たないため，AO-DCS は精度が低かった．また，属性数が大きくなるほど事例が属性空間上で疎に分布するため，NN-Ensemble の精度は低くなった．一方，提案手法は属性数が大きくなっても安定して高い精度を示すことが分かる．

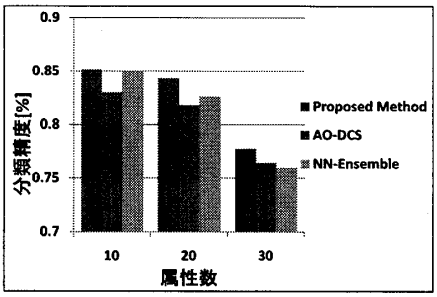

図 1: 属性数と分類精度の関係

4 結論

本論文では，Hoeffding Tree を用いた新たなアンサンブル学習法を提案した．人工的に作成したデータストリームを用いた実験により，提案手法の優位性を確認した．特に，属性数が多いほど提案手法による精度の改善は顕著であった．

参考文献

[1] Zhang, Y., Jin, X.: An automatic construction and organization strategy for ensemble learning on data streams. In: ACM SIGMOD Record. Volume 35. (2006) 28–33

[2] Zhu, X., Wu, X., Yang, Y.: Dynamic classifier selection for effective mining from noisy data streams. In: Proceedings of the 4th IEEE International Conference on Data Mining. (2004) 305–312

[3] Tsymbal, A., Pechenizkiy, M., Cunningham, P., Puuronen, S.: Dynamic integration of classifiers for handling concept drift. In: Journal of Information Fusion. Volume 9. (2008) 56–68

[4] Hulten, G., Spencer, L., Domingos, P.: Mining time-changing data stream. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2001) 97–106

[5] Gama, J., Medas, P., Rodrigues, P.: Learning decision trees from dynamic data streams. In: Proceedings of the 2005 ACM Symposium on Applied computing. (2005) 573–577