事例拡張を用いた半教師付き学習のデータストリームへの適用

Application of Instance Expansion to Semi-supervised Learning for Data Stream

小阪 達也 Tatsuya Kosaka　安村 禎明 Yoshiaki Yasumura　上原 邦昭 Kuniaki Uehara

神戸大学大学院工学研究科情報知能学専攻
Department of Computer Science and Systems Engineering, Graduate School of Engineering, Kobe University

In this report, we present a semi-supervised learning method for a data stream that contains labeled and unlabeled data. The instances in a data stream are assumed to arrive in a chunk. This method builds an ensemble of classifiers by semi-supervised learning per chunk. This method adds weighted instances in old chunks to a current chunk in order to supply a deficiency of labeled instances in the current chunk. The instances in the old chunk are weighted by TrAdaBoost algorithm.Concept change is detetected based on a sum of weights of diffencially classified instanes. The experiment using artificial data streams shows that the proposed method results higher accuracy.

1. はじめに

近年，データストリームから有用な規則やパターンを発見するマイニング技術の研究が盛んに行われている．データストリームとは，大量の電子化データの流れのことであり，従来のようにデータベースに蓄えられた大規模データを分析するのではなく，刻々と増え続けるデータの流れをリアルタイムに分析する必要がある．このようなデータストリームは，時間が経過すると，その性質が変化するという特徴を持つ．この特性は，concept change と呼ばれ，concept change の適切な処理はデータストリームマイニングを行う上での重要な課題となっている．

これまでデータストリームに対する学習手法が多く提案されてきた [?, ?, ?, ?]．これらの手法は，すべてのデータにラベルが付けられているストリームを対象として，concept change に対応している．しかしながら，実社会のデータストリームでは，ラベル付けに莫大なコストがかかるために，すべてのデータにラベルを付けることは現実的ではない．特にリアルタイムで，すべてのデータにラベル付けするのは困難である．

そこで本稿では，少量のデータのみにラベル付けされたストリームの分類問題を扱う．本手法では，得られたストリームデータを chunk と呼ばれる一定数の事例を含むデータ集合に分割し，1つの chunk 内のデータから1つの分類器を生成するアンサンブルアプローチ [?] をベースにする．このとき，少量のラベル付き事例を含む chunk から，いかにして精度の高い分類器を生成するかが問題となる．

本手法では，1つの chunk 内に含まれるラベル付き事例の不足による分類器の精度不足を軽減するために，過去の chunk 内に含まれるラベル付き事例を有効に利用する．また，本手法では，concept change に迅速に対応するために，change の検出を行う．この検出手法では，分類が困難な事例に対して小さな重みを付け，容易に分類できる事例に対しては大きな重みを付けることによって分類器による誤分類の影響を軽減する．この手法により concept change を正確に検出し，データストリームに対する分類精度の向上を目指す．

連絡先: 小阪 達也，神戸大学大学院工学研究科情報知能学専攻，神戸市灘区六甲台町 1-1，Tel: 078-803-6220，E-mail: t-kosaka@ai.cs.kobe-u.ac.jp

2. 半教師付き学習のデータストリームへの適用

本節では，半教師付きデータストリームに対する学習手法の詳細について述べる．

2.1 概要

図 **??**に本手法の全体の概要を示す．まず，ストリームデータを一定数の事例集合（chunk）に分割する．このとき chunk は，図 **??**で示すようにラベル付き事例とラベルなし事例によって構成されている．次に，この chunk 内のデータを現在の concept の主成分空間へと射影する前処理を施す．これは，データを分散がより大きな空間へ射影することで，データの分布の偏りを緩和するためである．また，ラベル付き事例の数を補完するために過去の chunk 内のラベル付き事例に TrAdaBoost[?] を適用して学習に有用な事例を選別して，新しい chunk のラベル付き事例に追加する．これらの前処理を施したデータに半教師付き学習手法の 1 つである Tri-training[?] を行って，1 つの分類器を生成する．最終的に，過去に生成した分類器の多数決で，新しい chunk に含まれる事例のクラスを予測する．過去に生成した分類器が新しい chunk の事例のクラスを正しく予測できないときに concept change を検出する．このとき，過去の分類器の削除と新しい分類器の追加を行うことで，concept change に対応する．また，新しい chunk のデータから新しい concept の主成分空間を導出する．

2.2 Tri-training

本手法では，図**??**における半教師付き学習として Tri-training[?] を用いる．Tri-training は，3 つの分類器を利用した co-training style（共訓練型）のアルゴリズムである．

Tri-training では，初めにラベル付き事例集合 L にブートストラップサンプリングを行って生成した事例集合 S_1，S_2，S_3 から 3 つの分類器 h_1，h_2，h_3 を生成する．任意の分類器 h_i に対して，h_i 以外の 2 つの分類器が同じラベルを予測したラベルなし事例に予測したラベルを付ける．例えば，h_2 と h_3 がラベルなし事例集合 U のある事例 $\boldsymbol{x}$ に対して同じラベルを予測したときは，その予測したラベルで事例 $\boldsymbol{x}$ にラベル付けをして，h_1 の再訓練に用いる．他の分類器についても同様のことを行う．この手続きを，3 つの分類器が更新されなくなるまで繰り返す．最後に，ラベルなし事例のラベルを 3 つの分類器

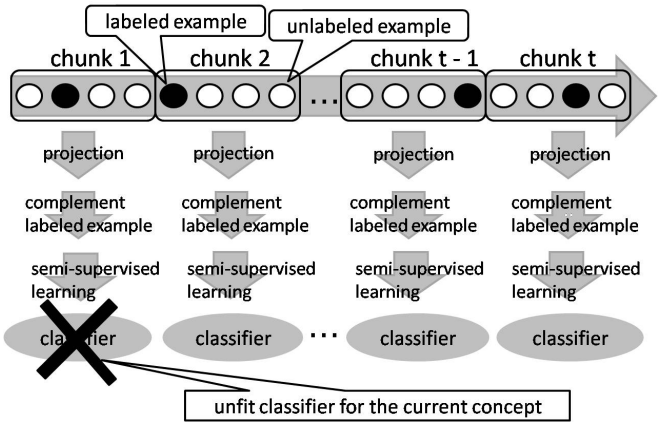

図 1: 概要

の多数決で決定する．

本手法では，Tri-training における初期分類器を生成する際にブートストラップサンプリングを用いることを避けるために，Tri らが提案した手法 [?] に従って，ラベル付き事例集合 L にそれぞれ異なる学習アルゴリズムを適用することで 3 つの異なる分類器 h_1，h_2，h_3 を生成する．

2.3 事例拡張による分類器の精度改善

図**??**のようにアンサンブルアプローチでは，ストリームから取得したデータを chunk と呼ばれる一定数（$ChunkSize$）の事例集合に分割し，1 つの chunk から 1 つの分類器を生成し，予測にはそれら複数の分類器の（重み付き）多数決をとる．しかしながら，半教師付きデータストリームを扱う場合は，1 つの chunk 内にラベル付き事例が少数しか含まれない．半教師付き学習では，ラベル付き事例を用いて学習した学習器でラベルなし事例のラベルを予測して擬似的なラベル付き事例として扱って訓練に利用するが，1 つの chunk 内のラベル付き事例の数が少なすぎるとラベルなし事例のラベルの予測の精度が悪化する．

この問題を解決するために，本研究では過去に取得した chunk 内のラベル付き事例を利用する．過去に取得した chunk 内のラベル付き事例を新しい chunk 内のラベルなし事例のラベルの予測に利用することで，ラベルなし事例のラベル予測の精度を改善し，半教師付き学習から得られる分類器の性能の安定を図る．しかしながら，単純に過去に取得した chunk 内のラベル付き事例を追加すると，concept change が起こった後の分類器の精度が悪化する．つまり，過去の chunk 内のラベル付き事例の中には，学習に有用な事例と有用でない事例が含まれている．そこで，本研究ではこれらの事例を区別するために TrAdaBoost[?] を利用する．

TrAdaBoost では，初めに対象ドメインのデータと対象ドメイン外のデータから学習した分類器を生成する．生成した分類器に対象ドメイン外のデータを分類させて，分類を誤った事例の重みを小さくし，正しく分類できた事例の重みは変化させない．また，生成した分類器に対象ドメインのデータを分類させて，分類を誤った事例の重みを大きくし，正しく分類できた事例の重みは変化させない．このように重み付けした両データで分類器を再訓練する．この手続きを繰り返すことで，対象ドメイン外のデータの中で対象ドメインのデータの分布に従うデータと従わないデータを区別する．

これを利用することで，過去の chunk 内のラベル付き事例の中で新しい chunk 内のラベル付き事例の分布に従う事例には大きな重みをつけ，新しい chunk 内のラベル付き事例の分布に従わない事例には小さな重みをつけることができる．したがって，concept が変化する前の事例の影響を軽減できる．

2.4 concept change の検出

concept change を含むデータへ迅速に適応するためには，モデルの大幅な修正が必要とされる大きな変化（concept shift）を発見することが重要である．従来のアンサンブル手法の多くは，穏やかな変化（concept drift）を対象にしており，明示的に変化が検出されることは少ない．そのため急激な変化が起こると，しばらく前回の concept を反映した分類器の影響が残り，学習の精度を低下させる原因となる．また，変化を検出することは，データの解析や統計処理においても役立つと考えられる．本稿では，以下のようにして concept change を検出し，迅速に変化に追従する．

2.1 で述べたように，過去の分類器が新しい chunk の事例を正しく分類できないときに concept change と判断するが，データの中には分類が困難な事例と分類が容易な事例が存在する．例えば，クラス境界付近の事例は分類器によって分類結果が変化しやすい．つまり，クラス境界付近の事例は，分類器の性能の影響を強く受けることになる．したがって，concept change の検出にはクラス境界付近でない事例に着目する．本手法では，この分類が困難な事例と分類が容易な事例の推定に Tri-training と TrAdaBoost を利用する．

2.4.1 Tri-training を用いた重み付け

Tri-training では，前述したようにラベルなし事例のラベルの予測は 3 つの分類器の多数決で決定される．本手法では，このときの多数決の結果を基に分類が容易な事例と分類が困難な事例を推定する．本研究では，2 クラス分類問題を扱う．このとき，ラベルなし事例のラベルの推定のために 3 つの分類器による多数決を行うと 2 通りの結果が得られる．つまり，全ての分類器が同じ分類結果である場合，2 つの分類器が同じ分類結果で残りの分類器が異なる分類結果である場合である．このとき，全ての分類器が同じ分類結果である場合は，推定したラベルなし事例のラベルに高い確信度を有する．したがって，そのラベルなし事例は分類が容易な事例と推定できる．

一方，2 つの分類器が同じ分類結果で残りの分類器が異なる分類結果である場合は，推定したラベルなし事例のラベルは確信度が低いと推察される．したがって，そのラベルなし事例は分類が困難な事例と推定できる．

このとき，事例 $(\boldsymbol{x}_i, y_i)$ に対する重みを $w(\boldsymbol{x}_i)$ で表し，Tri-training の多数決で用いる 3 つの分類器を $h_t(t = 1, \ldots, 3)$，分類が容易な事例に付与する重みを $w_{tri}(w_{tri} > 1)$ とすると，以下の式に従って事例の重みを更新する．

$$w(\boldsymbol{x_i}) = w(\boldsymbol{x_i}) \times \begin{cases} w_{tri} & \text{if } h_1(\boldsymbol{x_i}) = h_2(\boldsymbol{x_i}) = h_3(\boldsymbol{x_i}) \\ 1 & \text{otherwise.} \end{cases}$$

2.4.2 TrAdaBoost を用いた重み付け

本手法では，前述したように TrAdaBoost は新しい chunk のラベル付き事例の数を補うために用いられる．このとき，過去の chunk のラベル付き事例を追加する前と追加した後では学習した境界がわずかに変化する．本手法では，この境界の変化を基に分類が容易な事例と分類が困難な事例を推定する．つまり，新しい chunk に含まれるラベル付き事例のみから学習した分類器と新しい chunk に含まれるラベル付き事例に TrAdaBoost を適用して得た過去の chunk のラベル付き事例を追加して学習した分類器によるラベルなし事例の分類差に着目する．先ほど述べたように，過去の chunk のラベル付き事例を追加する

前と追加した後では境界がわずかに変化する．過去の chunk のラベル付き事例を追加する前と追加した後で分類結果が異なる事例は分類が困難な事例である可能性が高い．したがって，過去の chunk のラベル付き事例を追加する前と追加した後でも分類結果が同じ事例は分類が容易な事例である可能性が高い．このとき，事例 $(\boldsymbol{x}_i, y_i)$ に対する重みを $w(\boldsymbol{x}_i)$ で表し，新しい chunk のラベル付き事例から学習した分類器を C_A とする．新しい chunk のラベル付き事例に TrAdaBoost を適用して得た過去の chunk のラベル付き事例を追加して学習した分類器を C_B とし，分類が困難な事例に付与する重みを $w_{tra}(w_{tra} > 1)$ とすると，以下の式に従って事例の重みを更新する．

$$w(\boldsymbol{x_i}) = w(\boldsymbol{x_i}) \times \begin{cases} w_{tra} & \text{if } C_A(\boldsymbol{x_i}) = C_B(\boldsymbol{x_i}) \\ 1 & \text{otherwise.} \end{cases}$$

2.4.3 事例の重み付け

本手法では，以上の 2 つの考え方を組み合わせて分類が困難な事例と分類が容易な事例を推定する．つまり，ラベルなし事例は次の 4 種類に区別される．(i)Tri-training を用いた推定と TrAdaBoost を用いた推定の両方において分類が容易な事例であると推定された事例，(ii)Tri-training を用いた推定のみに分類が容易な事例であると推定された事例，(iii)TrAdaBoost を用いた推定のみに分類が容易な事例と推定された事例，(iv)Tri-training を用いた推定と TrAdaBoost を用いた推定の両方において分類が容易な事例であると推定されなかった事例である．したがって，以下の式で事例 $(\boldsymbol{x}_i, y_i)$ の重み $w(\boldsymbol{x}_i)$ を更新する．

$$w(\boldsymbol{x_i}) = w(\boldsymbol{x_i}) \times \begin{cases} w_{tri} w_{tra} \\ (\text{if } h_1(\boldsymbol{x_i}) = h_2(\boldsymbol{x_i}) = h_3(\boldsymbol{x_i})) \cap (\text{if } C_A(\boldsymbol{x_i}) = C_B(\boldsymbol{x_i})) \\ w_{tri} \\ (\text{if } h_1(\boldsymbol{x_i}) = h_2(\boldsymbol{x_i}) = h_3(\boldsymbol{x_i})) \cap (\text{if } C_A(\boldsymbol{x_i}) \neq C_B(\boldsymbol{x_i})) \\ w_{tra} \\ (\text{if } !(\text{if } h_1(\boldsymbol{x_i}) = h_2(\boldsymbol{x_i}) = h_3(\boldsymbol{x_i}))) \cap (\text{if } C_A(\boldsymbol{x_i}) = C_B(\boldsymbol{x_i})) \\ 1 \\ (\text{if } !(\text{if } h_1(\boldsymbol{x_i}) = h_2(\boldsymbol{x_i}) = h_3(\boldsymbol{x_i}))) \cap (\text{if } C_A(\boldsymbol{x_i}) \neq C_B(\boldsymbol{x_i})) \end{cases}$$

本手法では，このようにして事例に重みを付与し，過去の分類器と新しい分類器で分類差が生じた事例の重みの和で concept change を検出する．

3. 評価実験

本手法の有効性を評価するために人工データを用いて実験を行った．

3.1 実験設定

本手法の性能を評価するために，実験を行った．実験には，人工的に concept change を実現させた 2 種類の変動超平面データセットを用いる．

1 つ目は，2 次元空間の人工データである．各クラスの事例は正規分布に従うとする．positive クラスの事例は，平均 (m_1, n_1)，標準偏差 σ に従って生成し，negative クラスの事例は，平均 (m_2, n_2)，標準偏差 σ の正規分布に従って生成する．これらの生成した事例を z 軸を中心に θ 回転させる．この θ の値を変化させることで concept change を実現する．本実験では，2 つのクラスが平行になるように $m_1 = m_2$ とし，2 点間の距離が 1000 になるように n_1 と n_2 を決定している．また，標準偏差 $\sigma = 1000$ とした．以降，このデータをデータ A とする．

2 つ目は，d 次元空間の超平面 $\sum_{i=1}^{d} a_i x_i = a_0$ を用いる．各事例のラベル付けは $\sum_{i=1}^{d} a_i x_i \geq (a_0 + \alpha)$ のとき positive，$\sum_{i=1}^{d} a_i x_i \leq (a_0 - \alpha)$ のとき negative とする．このとき，$[a_0 - \alpha, a_0 + \alpha]^d$ に事例が存在しないように調整をした．また，本実験では $\alpha = 0.1$ とした．多次元空間 $[0, 1]^d$ に一様分布するランダム事例を生成し，等しい大きさの空間に分割する a_0 を $a_0 = \frac{1}{2}\sum_{i=1}^{d} a_i$ と設定する．このとき，重み a_i $(1 \leq i \leq d)$ を変化させることで concept change を実現する．本実験では，$d = 5$ とした．以降，このデータをデータ B とする．

なお，すべての実験において chunk 内の事例の数を $Chunksize = 200$，chunk 内のラベル付き事例の数を $|L| = 10$，chunk 内のラベルなし事例の数 $|U| = 190$ とする．このとき，chunk 内のラベル付き事例はランダムに選択される．ベース分類器には C4.5 を用いた．

3.2 実験１

まず，データの重みを利用した change の検出手法の有効性について評価する．実験には，2000 事例ごとに concept change を起こすデータ A とデータ B を用いた．また，実験においてデータに付与する重みはそれぞれ $w_{tra} = 3$，$w_{tri} = 3$ と設定する．また，ラベル付き事例の重み $w_L = 10$，ラベルなし事例の初期の重み $w_U = 1$ とした．

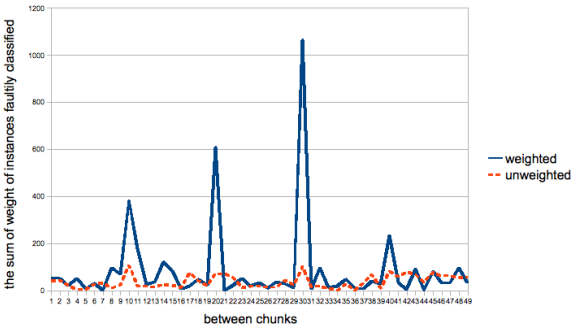

図 2: 分類差が生じた事例の重みの和の推移 (データ A)

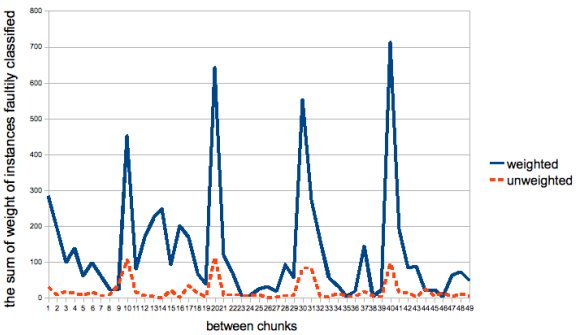

図 3: 分類差が生じた事例の重みの和の推移 (データ B)

3.3 実験１の結果と考察

データ A に対する分類差が生じた事例の重みの和の推移を図??，データ B に対する分類差が生じた事例の重みの和の推移を図??に示す．図??から，データ A において，重み付けしなかったときには検出できなかった change を，重みを利用することで検出が可能となっていることがわかる．また，図??

から，データ B において，重み付けを行わずとも change は検出が可能であったが，重みを利用することで検出がより容易になっていることがわかる．したがって，提案したデータの重みを利用した change の検出手法は有効であることがわかった．

しかしながら，データ A とデータ B の両方においても，全ての change を正確に検出できる場合と change の検出に遅延が発生することや誤った位置で検出する場合が見られた．これは，ラベル付き事例をランダムに選択しているため，分類器の精度が不安定であることが原因と考えられる．

3.4 実験２

次に，本稿が提案する事例拡張による分類器の精度改善手法の有効性を評価する．実験には，2000 事例ごとに concept change を起こすデータ A とデータ B を用いた．提案手法の有効性を評価するために，半教師付きデータストリームにおいて chunk ごとに過去のラベル付き事例に TrAdaBoost を適用して事例拡張を行って学習した分類器と事例拡張を行わずに学習した分類器を生成する．それぞれ生成した分類器の学習エラー率の平均を計算して，これをデータストリームに対する学習エラー率とする．この試行を 10 回繰り返し，10 個のデータストリームに対する学習エラー率の平均を計算して比較した．

	事例拡張あり	事例拡張なし
データ A	3.91	4.10
データ B	7.72	9.53

表 1: 平均学習エラー率

3.5 実験２の結果と考察

実験結果を表**??**に示す．表**??**から，データ A とデータ B の両方において分類器の精度が改善していることがわかる．このことから，事例拡張による分類器の精度改善手法は有効であることがわかった．

しかし，対象となるデータによって精度の改善の大きさが異なることから，この手法の効果は対象とするデータによって差が生じることが考えられる．

4. まとめ

本稿では，半教師付きデータストリームに対する学習手法を提案した．半教師付き学習に事例拡張を行うことで，全体的な分類精度が向上した．また，分類の困難さに基づいて事例に付与した重みを利用することで，より容易に concept change の検出が可能となった．今後の課題としては，対象とするデータを検討し，それに相応しいラベル付き事例の選択方法を考慮する必要がある．

謝辞

この研究は，栢森情報科学振興財団の助成を受けて遂行された．

参考文献

[Klinkenberg 00] Klinkenberg, R., Joachims, T. : Detecting concept drift with support vector machines, Proc. of the 17th Int. Conf. on Machine Learning, pp. 487-494 (2000).

[Street 01] Street, W.N., Kim, Y. : A Streaming Ensemble Algorithm (SEA) for Large-Scale Classification, Proc. of the 7th ACM Int. Conf. on Knowledge Discovery and Data Mining, pp. 377-382 (2001).

[Kolter 05] Kolter, J., Maloof, M. : Using Additive Expert Ensembles to Cope with Concpet Drift, Proc. of the 22nd Int. Conf. on Machine Learning, pp. 449-456 (2005).

[Wang 03] Wang, H., Fan, W., Yu, P. S. and Han, J. : Mining concept-drifting data streams using ensemble classifiers, Proc. of the 9th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining, pp.226-235 (2003).

[Zhou 05] Z.H. Zhou. and Ming, L. : Tri-training: Exploiting Unlabeled Data Using Three Classifiers, IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 11, pp. 1529-1541 (2005).

[Tri 08] Tri, N.T., Le, N.M. and Shimazu, A. : Using Semi-supervised Learning for Question Classification, Information and Media Technologies, Vol. 3, No. 1, pp. 112-130 (2008).

[Dai 07] Dai, W., Yang, Q., Xue, G.R. and Yu, Y. : Boosting for transfer learning, ICML '07: Proc. of the 24th int. conference on Machine learning, pp. 193-200 (2007).

[Yasumura 07] Yoshiaki Yasumura, Naho Kitani and Kuniaki Uehara. : Quick Adaptation to Changing Concepts by Sensitive Detection, Proc. of the 20th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems, pp. 855-864 (2007).