

二つのオンラインクラス分類器を用いた Concept Drift の学習と検出

Learning and Detecting Concept Drift with Two Online Classifiers

西田 京介*1*2 山内 康一郎*2
 Kyosuke Nishida Koichiro Yamauchi

*1 日本学術振興会特別研究員
 JSPS Research Fellow

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

There are serious problems with online learning, especially in environments where the statistical properties of the target variable change over time. This change, known as *concept drift*, can happen either gradually or suddenly and significantly. Detecting concept drift is important for dealing with real-world online learning problems. We previously proposed a method of accurately detecting concept drift, the STEPDP method, but its false alarms degrade the predictive accuracy because STEPDP uses only one online classifier. To reduce the bad influence of false alarms on predictive accuracy, we have proposed a system that uses two online classifiers, which we call the Todi system. Using two online classifiers also enables Todi to accurately notify a user of the occurrence of concept drift. We demonstrate that Todi performed well with a spam filtering dataset.

1. はじめに

機械学習の研究において、順次与えられるサンプルを追加的に学習するオンライン学習、特に、学習対象の基となる統計的な性質が時間と共に変化する環境下での学習については多くの課題が残されている。この変化は *concept drift* と呼ばれ、緩やかな変化のみならず、突然かつ大きな変化をも含む [Tsymbal 04]。変化の性質によって学習システムが採るべき戦略は大きく異なるため、様々な性質の変化を含む問題の解決は非常に難しい。本研究では、入力サンプルに対応するクラスを予測するクラス分類問題においてこの解決を目指す。

2. Todi システム

我々は以前に、concept drift の検出法として STEPDP を提案した [Nishida 07]。STEPDP は単一のオンラインクラス分類器についてその予測精度を監視し、concept drift の発生が引き起こす最近の予測精度の急激な悪化を統計的検定によって検出する。STEPDP は他の従来手法に比べ高い検出性能を実現したが、検出時に単一のクラス分類器を初期化するため、誤検出が発生した場合にシステムの予測精度が大きく悪化する問題があった。そこで、誤検出が発生しても予測精度が悪化しない学習システムの実現を目指して、二つのオンラインクラス分類器を利用して concept drift の学習と検出を行うシステム Todi (two online classifiers system for learning and detecting concept drift) を我々は提案した [Nishida 08]。Todi は、変化検出時に初期化されるクラス分類器に加えて、検出後も継続して学習を行うクラス分類器を使用する。

2.1 クラス分類器の学習と選択

Todi は時刻 t に与えられた新たな学習サンプル (\mathbf{x}_t, y_t) を二つのオンラインクラス分類器 H_0 と H_1 で学習させ、その後それぞれの分類結果 $w_{0,t}$ と $w_{1,t}$ をサイズ W の時間窓に記憶する。さらに、Todi は五つのカウンタ n , r_0 , r_1 , s_0 及び s_1 を管理する。ここで、 n は H_0 が学習した全サンプル数; s_i は最近 W サンプルにおける H_i の正解数; r_i は $n - W$ サンプ

ル (全 n サンプルから最近 W サンプルを除いたもの) における H_i の正解数を表す。ここで、 H_i には任意のオンラインクラス分類器が使用できる。

次に、Todi はシステム出力を決定するために、現在の学習対象により適合したクラス分類器の出力を選択する。すなわち、もし $s_0 > s_1 + 1$ が真ならば $H_0(\mathbf{x}_t)$ を選択し、 $s_0 < s_1 - 1$ が真ならば $H_1(\mathbf{x}_t)$ を選択する。それ以外の場合は前時刻 $t - 1$ で選択されたクラス分類器の出力 $H_c(\mathbf{x}_t)$ を選択する ($c = 0$ or 1)。また、学習開始時には H_0 を選択し、concept drift の発生直後は H_1 を選択する。この選択方法は、システム出力が H_0 と H_1 の間で頻繁に切り替わり、その結果として予測精度を低下させることを防ぐ。

2.2 Concept Drift の検出

Todi は二群の母比率の差の検定により concept drift を検出する。 H_0 の過去の予測精度と最近の予測精度を $\hat{p}_A = r_A/n_A = r_0/(n - W)$ と $\hat{p}_B = r_B/n_B = s_0/W$ として定義したとき、片側の有意確率 (P_d) を得るために、以下の統計量 T を標準正規分布のパーセント点と比較する ($\hat{p} = (r_A + r_B)/(n_A + n_B)$)。

$$T(r_A, r_B, n_A, n_B) = \frac{|\hat{p}_A - \hat{p}_B| - 0.5(1/n_A + 1/n_B)}{\sqrt{\hat{p}(1 - \hat{p})(1/n_A + 1/n_B)}} \quad (1)$$

α を有意水準としたとき、 $P_d < \alpha$ かつ $\hat{p}_A > \hat{p}_B$ が真ならば帰無仮説 (母比率 p_A と p_B が等しい; concept drift は発生していない) が棄却され、対立仮説 ($p_A > p_B$; concept drift が発生した) が採択される。すなわち、Todi は concept drift が引き起こす H_0 の最近の予測精度 $\hat{p}_B = s_0/W$ の急激な低下を検出する。なお、 $n \geq 2W$ を満たすまで検出は行わない。

2.3 Concept Drift 発生の通知

追加したオンラインクラス分類器 H_1 を使用することで、前回の検出 (時刻 $t - n$) において真に concept drift が発生していたかどうかを検証可能になる。Todi は concept drift を検出した後、統計量 $T(r_0 + s_0, r_1 + s_1, n, n)$ の片側有意確率 P_c を検出時と同様に計算し、もし $P_c < \beta$ かつ $(r_0 + s_0)/n > (r_1 + s_1)/n$ が真ならば前回の検出が正しかったとみなし、システムの利用者に concept drift の発生を通知する。ここで、 β はもう一つの有意水準である。すなわち、前回の H_0 の初期化が有意に有効であった場合にユーザは concept drift の発生を通知される。

連絡先: 西田京介, 060-0814 札幌市北区北 14 条西 9 丁目,
 011-706-6815, knishida@complex.eng.hokudai.ac.jp

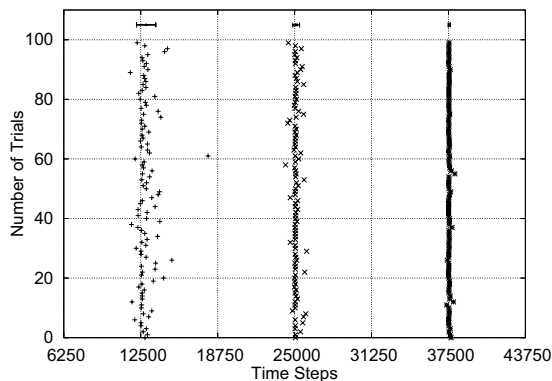


図 1: Results that Todi notified a user of the occurrence of concept drift on the SEA concepts. These results were clustered by the Weka implementation of EM algorithm (`numClusters = 3`) [Witten 05]. Error bars show the standard deviations of each cluster from 100 trials.

もし $P_c < \beta$ が真ならば (前回の検出が正しかった場合), Todi は H_0 (s_0 と $\{w_{0,t}\}$ を含む) を H_1 にコピーして保持した後, H_0 を初期化する. 一方, $P_c < \beta$ が偽ならば H_1 を保持し, H_0 は初期化する. なお, n, r_0, r_1, s_0 及び $\{w_{0,t}\}$ は, $P_d < \alpha$ が真のとき $P_c < \beta$ の真偽に関わらず初期化する. この処理によって, 変化検出前の学習対象により適合していたクラス分類器が H_1 に残される.

3. 計算機実験

まず, Todi の変化検出の正確さについて検証するため, concept drift の評価用によく用いられるデータセットである SEA concepts [Street 01] を使用して実験を行った. 入力 \mathbf{x} は $\mathbf{x} \in [0, 10]^3$ で与えられ, クラス y は $x_0 + x_1 \leq \theta$ が真のとき 1, 偽のとき 0 となる. この閾値 θ が, 8, 9, 7, そして 9.5 の順に 12500 サンプル毎に変化する. なお, 閾値の変化量が最も小さい (8 \rightarrow 9) とき, すなわち時刻 12500 における変化の検出が最も難しい. ノイズは 10% のサンプルのクラスを反転させて導入した. また, Todi が内部で用いるオンラインクラス分類器には連続値を処理可能な Naive Bayes を使用した [John 95, Equations (1)–(3)]. 図 1 に Todi が使用者に concept drift の発生を通知した結果について示す. Todi はノイズが多く含まれる環境下でも, 二つのクラス分類器を使用して変化検出の検証を行うことで, 真の concept drift のみをユーザに通知することができた.

次に, Todi の学習システムとしての性能を検証するため, TREC 2006 Spam Track Public Corpora [Cormack 06] を使用して実験を行った. このデータセットには時系列順に並べられた 37,822 通のメールが含まれる. スпамフィルタリング課題においては, 受信メール内容の変化や狡猾なスパム送信者が起こす変化など, 日々発生する様々な性質の変化に対応しながら学習を進めなければならない. Todi と STEPDP が内部で用いるオンラインクラス分類器にはベイジアンフィルタ Bogofilter [Raymond 07] を使用した. 図 2 に Todi, STEPDP, Bogofilter の 1000 メールごとの誤分類率を示す. STEPDP は誤検出によって分類精度が度々大きく悪化したが, Todi は安定した学習を実現できた. また, Todi が真の concept drift が発生したと確認した後の期間は単体の Bogofilter に比べて特に低い誤分類率を達成できた.

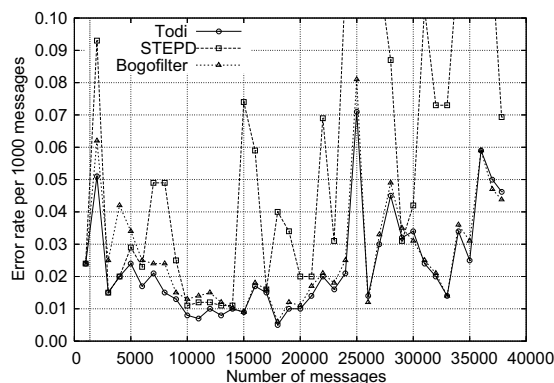


図 2: Error rates per 1000 messages of Todi, STEPDP, and Bogofilter on the TREC 2006 Spam Track Public Corpora (English, trec06p/full, immediate). The vertical line indicates the concept drift that Todi notified to a user.

4. むすび

本研究では, 二つのオンラインクラス分類器を利用して concept drift の学習と検出を行うシステム Todi を提案した. concept drift が含まれる人工データセットにおいて Todi が concept drift を正しくユーザに通知できることを確認した. また, 実データであるスパムフィルタリング課題において, Todi は単体のスパムフィルタ Bogofilter より高い性能を実現した.

参考文献

- [Cormack 06] Cormack, G. V.: TREC 2006 Spam Track Public Corpora (2006)
- [John 95] John, G. H. and Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers, in *Proc. 11th Conf. Uncertainty in Artificial Intelligence*, pp. 338–345 (1995)
- [Nishida 07] Nishida, K. and Yamauchi, K.: Detecting Concept Drift using Statistical Testing, in *Proc. 10th Int. Conf. Discovery Science*, pp. 264–269 (2007)
- [Nishida 08] Nishida, K.: *Learning and Detecting Concept Drift*, Dissertation, Hokkaido University (2008), [Online]. Available: <http://kussharo.complex.eng.hokudai.ac.jp/~knishida/>
- [Raymond 07] Raymond, E. S., Relson, D., Andree, M., and Louis, G.: *Bogofilter* (2007), v1.1.5
- [Street 01] Street, W. N. and Kim, Y. S.: A streaming ensemble algorithm (SEA) for large-scale classification, in *Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 377–382 (2001)
- [Tsymbal 04] Tsymbal, A.: The problem of concept drift: definitions and related work, Technical Report TCD-CS-2004-15, Department of Computer Science, Trinity College Dublin (2004)
- [Witten 05] Witten, I. H. and Frank, E.: *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, second edition (2005)