Adaptive Support Vector Machine for Time-Varying Data Streams Using Martingale

Shen-Shyang Ho and Harry Wechsler
George Mason University
Department of Computer Science
4400 University Drive, Fairfax, VA 22030
{sho, wechsler}@gmu.edu

Abstract

A martingale framework is proposed to enable support vector machine (SVM) to adapt to time-varying data streams. The adaptive SVM is a one-pass incremental algorithm that (i) does not require a sliding window on the data stream, (ii) does not require monitoring the performance of the classifier as data points are streaming, and (iii) works well for high dimensional, multi-class data streams. Our experiments show that the novel adaptive SVM is effective at handling time-varying data streams simulated using both a synthetic dataset and a multi-class real dataset.

1 Introduction

In this paper we propose an efficient adaptive support vector machine (SVM) for time-varying data streams based on the martingale approach [Vovk *et al.*, 2003] and using adiabatic incremental learning [Cauwenberghs and Poggio, 2000]. When a new data point is observed, hypothesis testing decides whether any change has occurred. Once a change is detected, historical information about previous data is removed from the memory.

2 Change Detection Using Martingale

[Vovk *et al.*, 2003] introduced the idea of testing exchangeability online using martingale. The sequence of random variables $z_1, \cdots, z_n$ is exchangeable if the joint distribution $p(z_1, \cdots, z_n)$ is invariant under any permutation of the indices of the random variables. When a new labeled example, z_n, is observed, testing exchangeability of the sequence of examples $z_1, z_2, \cdots, z_n$ consists of two main steps [Vovk *et al.*, 2003]:

1. Extract the randomized p-value for z_n:
$$p_n = \frac{\#\{i : \alpha_i > \alpha_n\} + \theta_n \#\{i : \alpha_i = \alpha_n\}}{n} \quad (1)$$
where α_i is the strangeness measure (for SVM, one can use either the Lagrange multipliers or the distances from the hyperplane) for z_i, $i = 1, 2, \cdots, n$ and θ_n is a random number between 0 and 1.

2. Construct the randomized martingale:
$$M_n^{(\epsilon)} = \prod_{i=1}^{n} \left(\epsilon p_i^{\epsilon-1}\right) \quad (2)$$
where p_i are p-values, a constant $\epsilon \in [0, 1]$, and the initial martingale $M_1^{(\epsilon)} = 1$.

p_i and $M_i^{(\epsilon)}$, for $i = 2, \cdots, n-1$, are computed using Step 1 and 2.

The exchangeability of a data sequence is a slightly weaker assumption compared to the randomness (i.i.d.) assumption. Hypothesis testing using the randomized martingale is suitable to check for the randomness of a data sequence which, in turn, can detect changes.

One tests the null hypothesis H_0 : "no change in the data stream" against the alternative hypothesis H_1 : "change in the data stream". The martingale test continues as long as
$$0 < M_n^{(\epsilon)} < \lambda \quad (3)$$
One rejects the null hypothesis when $M_n^{(\epsilon)} \geq \lambda$.

The martingale test is justified using the Doob's Maximal Inequality [Steele, 2001] which states that for a martingale $\{M_k\}$ and any $\lambda > 0$,
$$\lambda P\left(\max_{k \leq n} M_k \geq \lambda\right) \leq E(M_n) \quad (4)$$
Hence, if $E(M_n) = E(M_1) = 1$, then
$$P\left(\max_{k \leq n} M_k \geq \lambda\right) \leq \frac{1}{\lambda} \quad (5)$$
This inequality means that it is unlikely for M_n to have a high value. Based on this inequality, λ is appropriately set to reject the null hypothesis.

3 Experiments

Experiments are first conducted on a binary-class two-dimensional rotating hyperplane data stream [Hulten *et al.*, 2001]. To show the performance of the adaptive SVM on high dimensional multi-class data stream, experiments are carried out using a modified three-digit data stream based on the USPS handwritten digit dataset [LeCun *et al.*, 1989]. For the SVMs, $C = 10$ and the Gaussian kernel are used.

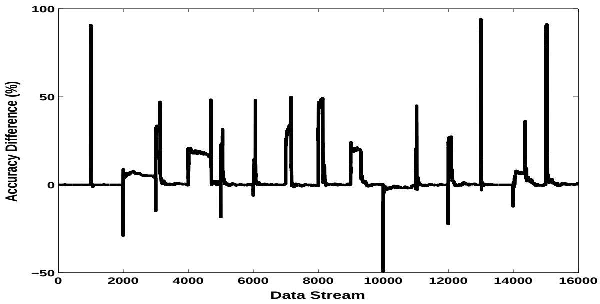

Figure 1: The difference between baseline accuracy and the accuracy of the adaptive SVM. Negative values imply better performance than the baseline accuracy.

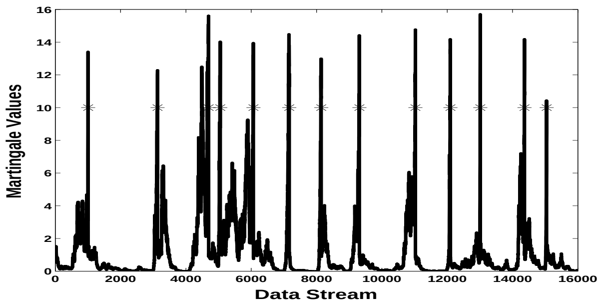

Figure 2: The martingale values of the data stream. Detected change points (denoted $*$): 1007, 3133, 4697, 5054, 6068, 7165, 8144, 9312, 11029, 12099, 13015, 14368, 15045. Miss detections: 2001, 10001. The median of the time between true change points and their corresponding detected change points is 99 time units.

A. 2-D Rotating Hyperplane Synthetic Dataset

For the two-dimensional rotating hyperplane data stream, the sequence consists of 16,000 data points with changes in the data distribution occurring at points $(1000 \times i) + 1$, for $i = 1, 2, \cdots, 15$. For each segment of a steady data distribution, 500 points from the same data distribution are used to assess the performance of the SVM classifier.

The adaptive SVMs using $\lambda = 16$, for the martingale test (3), displays very good performance as shown in Figure 1. The difference in accuracy from baseline accuracy (performance of the classifier that adapts to the given true change points) is near zero most of the time. Short bursts of accuracy discrepancy occur between true change points and their corresponding detected change points. Longer accuracy discrepancies happen when a significant change is detected late (e.g. change point 4001) as shown in Figure 1. In Figure 2, we observe that the martingale value is small and stable when no change in the data distribution is detected. The martingale value increases abruptly when change is suspected.

B. Three-digit Data Stream

The USPS handwritten digits dataset, which consists of 10 classes of dimension 256 and includes 7,291 data points, is modified to form a data stream as follows. There are four different data segments. Each segment draws from a fixed set of three different digits in a random fashion. The three-digit set changes from one segment to the next. The composition of the data stream and ground truth for the change points are summarized in Table 1. Ten simulations were performed on the simulated data stream using the one-against-the-rest multi-class SVM and the martingale test with $\lambda = 10$.

Segment	Digit 1	Digit 2	Digit 3	Total	Change Point
1	597/359 (0)	502/264 (1)	731/198 (2)	1830/821	1831
2	597/359 (0)	658/166 (3)	652/200 (4)	1907/725	3738
3	503/264 (1)	556/160 (5)	664/170 (6)	1723/594	5461
4	645/147 (7)	542/166 (8)	644/177 (9)	1831/490	-

Table 1: **Three-Digit Data Stream**: TR/TS (D): For the true digit class D, TR and TS are the number of data points used for segment construction and for testing respectively.

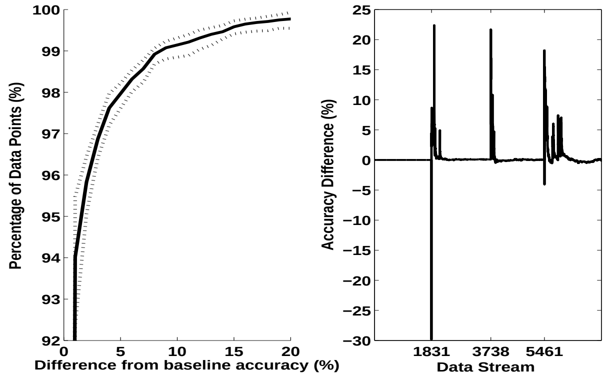

Figure 3: Left Graph: The proportion of data points with the difference from the baseline accuracy less than or equal to some percentage value. The maximum, minimum and mean percentage of data points at each accuracy difference are shown. Right Graph: The mean difference between the baseline accuracy and the accuracy of the adaptive SVM.

For the three-class data stream there are three martingale values that have to be computed at each point in order to detect change. When one of the martingale values is greater than λ, the SVM will detect and adapt to the change. One can see from Figure 3 that the adaptive SVM performs extremely well. About 94% of the data points have a difference from baseline accuracy less than or equal to 1%. There are no miss detections for all the ten simulations.

References

[Cauwenberghs and Poggio, 2000] G. Cauwenberghs and T. Poggio. Incremental support vector machine learning. In *Advances in Neural Information Processing Systems 13*, pages 409–415. MIT Press, 2000.

[Hulten *et al.*, 2001] Geoff Hulten, Laurie Spencer, and Pedro Domingos. Mining time-changing data streams. In *KDD*, pages 97–106, 2001.

[LeCun *et al.*, 1989] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. J. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989.

[Steele, 2001] M. Steele. *Stochastic Calculus and Financial Applications*. Springer Verlag, 2001.

[Vovk *et al.*, 2003] Vladimir Vovk, Ilia Nouretdinov, and Alexander Gammerman. Testing exchangeability on-line. In Tom Fawcett and Nina Mishra, editors, *ICML*, pages 768–775. AAAI Press, 2003.