Online Imbalanced Learning with Kernels

Haiqin Yang, Junjie Hu, Michael R. Lyu, and Irwin King
Shenzhen Key Laboratory of Rich Media Big Data Analytics and Applications
Shenzhen Research Institute, The Chinese University of Hong Kong
Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
{hqyang, jjhu, lyu, king}@cse.cuhk.edu.hk

Abstract

Imbalanced learning, or learning from imbalanced data, is a challenging problem in both academy and industry. Nowadays, the streaming imbalanced data become popular and trigger the volume, velocity, and variety issues of learning from these data. To tackle these issues, online learning algorithms are proposed to learn a linear classifier via maximizing the AUC score. However, the developed linear classifiers ignore the learning power of kernels. In this paper, we therefore propose online imbalanced learning with kernels (OILK) to exploit the non-linearity and heterogeneity embedded in the imbalanced data. Different from previously proposed work, we optimize the AUC score to learn a non-linear representation via the kernel trick. To relieve the computational and storing cost, we also investigate different buffer update policies, including first-in-first-out (FIFO) and reservoir sampling (RS), to maintain a fixed budgeted buffer on the number of support vectors. We demonstrate the properties of our proposed OILK through detailed experiments.

1 Introduction

Streaming imbalanced data become more and more popular in various real-world applications, such as abnormal behaviors in surveillance systems, fraudulence in credit card transactions, and clicking/browsing behaviors in online ads/news. In these applications, the interesting events are usually important, but rarely appear, which belongs to a minority class, while most other events are common and not so interested, which can be deemed as a majority class. The streaming characteristic of the imbalanced data triggers the problem of volume and velocity. Moreover, the variety of the data also increases the difficulty of learning from them when data appear sequentially.

To tackle the above problems, recent development of online learning algorithms have been proposed to learn a linear classifier via maximizing the Area Under the receiver operating characteristic curve (AUC) [2, 10]. The work also ignites the investigation of deriving theoretical generalization bound for pairwise loss functions [4, 9]. A main insufficiency of previously proposed is that they ignore the learning power of kernel methods and its good performance in online learning setting [5, 6].

To compensate this insufficiency, in this paper, we investigate online imbalanced learning with kernels (OILK) to exploit the non-linearity and heterogeneity of the imbalanced data, which is untouched yet. Our work is different from previously proposed online learning classifiers for cost-sensitive learning [2, 10], which only considers the model in a linear form. Meanwhile, our work is also different from NORMA [5], which aims at optimization the classification accuracy instead of the AUC score. Moreover, we adopt different strategy to conquer the computational and storing cost of online learning with kernels. That is, the number of support vectors can be scaled with the number of samples appeared [5]. Different from the truncation method adopted in [5], we maintain a buffer with a fixed budget to store the informative support vectors and adopt oblivious strategy to update the support vectors in the buffer when it is full. More specifically, two effective buffer update

policies, first-in-first-out (FIFO) and reservoir sampling (RS), are investigated. We conduct detailed experiments in various benchmark cost-sensitive learning datasets to demonstrate the properties of the proposed OILK model.

2 Online Imbalanced Learning with Kernels (OILK)

2.1 Problem Definition

We aim at learning a non-linear classifier for a binary classification problem with imbalanced data distributions for the two classes. Suppose the instance space is $\mathcal{X} \in \mathbb{R}^d$ and the label set is $\mathcal{Y} = \{-1, +1\}$. Let $\mathcal{P}$ denote an unknown (underlying) distribution over $\mathcal{X} \times \mathcal{Y}$ and the t-th sample, $\mathbf{z}_t = (\mathbf{x}_t, y_t)$, is drawn identically and independently from the distribution $\mathcal{P}$, where the t-th sample is $\mathbf{z}_t = (\mathbf{x}_t, y_t)$, $\mathbf{x}_t \in \mathcal{X}$ and $y_t \in \mathcal{Y}$. Without loss of generality, we assume the streaming data is unbalanced and the positive class is the minority class.

Following the same setup of NORMA [5], we assume the class of the learned non-linear decision functions, $f : \mathcal{X} \to \mathcal{Y}$ are elements of a Reproducing Kernel Hilbert Space (RKHS), $\mathcal{H}$. That is, there exists a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and a dot product $\langle \cdot, \cdot \rangle$ such that 1) it satisfies the reproducing property, $\langle k(\mathbf{x}, \cdot), f(\cdot) \rangle = f(\mathbf{x})$; 2) $\mathcal{H}$ is the closure of the span of all $k(\mathbf{x}, \cdot)$ with $x \in \mathcal{X}$. In other words, all $f \in \mathcal{H}$ are linear combinations of kernel functions.

Different from typical empirical loss for standard classification problems, we consider AUC metric [1, 2, 3, 10] in the following. Suppose the training data are separated into two sets, a positive dataset, $\mathcal{D}^+ = \{\mathbf{z}_i^+ = (\mathbf{x}_i^+, y_i) \in \mathbb{R}^d \times \{+1\}\}_{i=1}^{|\mathcal{D}^+|}$, and a negative dataset, $\mathcal{D}^- = \{\mathbf{z}_j^- = (\mathbf{x}_j^-, y_j) \in \mathbb{R}^d \times \{-1\}\}_{j=1}^{|\mathcal{D}^-|}$, the AUC score of the function f on these two sets, $\mathcal{D}^+$ and $\mathcal{D}^-$, is defined as

$$\text{AUC}(f) = \frac{\sum_{i=1}^{|\mathcal{D}^+|} \sum_{j=1}^{|\mathcal{D}^-|} \mathbb{I}[f(\mathbf{x}_i^+) - f(\mathbf{x}_j^-) > 0]}{|\mathcal{D}^+||\mathcal{D}^-|} = 1 - \frac{\sum_{i=1}^{|\mathcal{D}^+|} \sum_{j=1}^{|\mathcal{D}^-|} \mathbb{I}[f(\mathbf{x}_i^+) - f(\mathbf{x}_j^-) \leq 0]}{|\mathcal{D}^+||\mathcal{D}^-|} \tag{1}$$

where $\mathbb{I}[\cdot]$ is the indicator function which outputs 1 if the argument is true and 0 otherwise. Hence, maximizing AUC(f) is equivalent to minimizing $\sum_{i=1}^{|\mathcal{D}^+|} \sum_{j=1}^{|\mathcal{D}^-|} \mathbb{I}[f(\mathbf{x}_i^+) - f(\mathbf{x}_j^-) \leq 0]$. As directly optimizing AUC is equivalent to a combinatorial optimization problem, which is an NP-hard problem, we replace the indicator function by its convex surrogate, i.e., the hinge loss function

$$\ell_h(f, \mathbf{z}, \mathbf{z}') = \max(0, 1 - \frac{1}{2}(y - y')(f(\mathbf{x}) - f(\mathbf{x}'))), \tag{2}$$

and find the optimal decision function by minimizing the following objective

$$\mathcal{L}(f) = \frac{1}{2}\|f\|_{\mathcal{H}}^2 + C \sum_{i=1}^{|\mathcal{D}^+|} \sum_{j=1}^{|\mathcal{D}^-|} \ell_h(f, \mathbf{z}_i^+, \mathbf{z}_j^-), \tag{3}$$

where $\frac{1}{2}\|f\|_{\mathcal{H}}^2$ is the regularization term to control the functional complexity, $C > 0$ is a positive parameter quantifying the tradeoff of the regularization term and the error.

2.2 OILK for AUC Maximization

Our objective is to develop an online learning algorithm to efficiently update the non-linear decision function in (3). Taking into account the pairwise loss on the AUC approximation, we can define the *instantaneous regularized risk of AUC* on a single sample $\mathbf{z}_t$, by

$$\mathcal{L}(f, \mathbf{z}_t) = \frac{1}{2}\|f\|_{\mathcal{H}}^2 + C \sum_{i=1}^{t-1} \ell_h(f, \mathbf{z}_t, \mathbf{z}_i) \tag{4}$$

It should be noted that: 1) Based on the loss defined in (2), the t-th sample, $\mathbf{z}_t$, will not yield loss for previous samples with the same label; 2) The expression of (4) is different from the regularized risk of NORMA in [5] as it introduces the pairwise loss on all previously coming samples; 3) As in (4), if we need to store all previous samples, it is intractable for large-scale applications.

To resolve the computation and storing issues of (4), we decide to maintain a buffer, B_t, with a fixed budget, N, to store the most informative samples, which are also support vectors for the decision function, at time t. We then define the *budgeted instantaneous regularized risk of AUC* as follows

$$\mathcal{L}_{B_t}(f, \mathbf{z}_t) = \frac{1}{2}\|f\|_{\mathcal{H}}^2 + C\sum_{i=1}^{|B_t|} \ell_h(f, \mathbf{z}_t, \mathbf{z}_i). \tag{5}$$

In the above, the regularization is kept for the sake of preventing the hypothesis not moving too far in one direction when a change occurs.

Hence, we can perform classical stochastic gradient descent on the budgeted instantaneous regularized risk of AUC to update the decision function by

$$f_{t+1} := f_t - \eta \partial_f \mathcal{L}_{B_t}(f, \mathbf{z}_t)|_{f=f_t} \tag{6}$$

where $\eta > 0$ is the learning rate, which can be constant or decrease as the number of trials increases.

To evaluate the gradient of $\mathcal{L}_{B_t}$ with respect to f , we first calculate the (sub)gradient of ℓ_h with respect to f

$$\partial_f \ell_h(f, \mathbf{z}_t, \mathbf{z}_i) = \begin{cases} 0, & \phi(\mathbf{z}_t, \mathbf{z}_i) \geq 1 \vee y_t = y_i \\ -\varphi(\mathbf{z}_t, \mathbf{z}_i), & \phi(\mathbf{z}_t, \mathbf{z}_i) < 1 \wedge y_t \neq y_i \end{cases} \tag{7}$$

where $\phi(\mathbf{z}_t, \mathbf{z}_i) = \frac{1}{2}(y_t - y_i)(f(\mathbf{x}_t) - f(\mathbf{x}_i))$ and $\varphi(\mathbf{z}_t, \mathbf{z}_i) = \frac{1}{2}(y_t - y_i)(k(\mathbf{x}_t, \cdot) - k(\mathbf{x}_i, \cdot))$.

Hence, by substituting (7) into $\partial \mathcal{L}_{B_t}(f_t, \mathbf{z}_t)$, we obtain

$$\partial \mathcal{L}_{B_t} = f_t - C\sum_{i=1}^{|B_t|} \mathbb{I}[\phi(\mathbf{z}_t, \mathbf{z}_i) < 1 \wedge y_t \neq y_i] \cdot \varphi(\mathbf{z}_t, \mathbf{z}_i) \tag{8}$$

Now, we choose a zero initial hypothesis $f_1 = 0$ and express the decision function at t-th iteration as a kernel expansion while updating the $(t+1)$-th iteration in an incremental mode,

$$f_t(\mathbf{x}) = \sum_{i=1}^{|B_t|} \alpha_i k(\mathbf{x}_i, \mathbf{x}), \qquad f_{t+1}(\mathbf{x}) = f_t(\mathbf{x}) + \alpha_t k(\mathbf{x}_t, \mathbf{x}) \tag{9}$$

We can then derive the updating rule for the coefficients as follows

$$\alpha_i = \begin{cases} \eta C \sum_{i=1}^{|B_t|} \mathbb{I}[\phi(\mathbf{z}_t, \mathbf{z}_i) < 1 \wedge y_t \neq y_i] y_t, & i = t \\ (1-\eta)\alpha_i - \frac{\eta C}{2}\mathbb{I}[\phi(\mathbf{z}_t, \mathbf{z}_i) < 1 \wedge y_t \neq y_i](y_t - y_i), & i \neq t \end{cases} \tag{10}$$

It is worth to emphasize several remarks on the updating rule: 1) When a new sample does not incur errors or the label of previously stored support vectors is the same as that of the new sample, the updating rule is the same as NORMA [5]; 2) When the new sample introduces errors, i.e., the pairwise distances of the new sample and the support vectors with opposite labels are too close, the updating rule is different from NORMA, which only performs one way updating when a new sample appears. However, our OILK works intuitively and is especially in favor of imbalanced data as it keeps the balance of updating: the magnitude of its coefficient at the new sample is proportional to the count sum of pairwise mistakes made on previously stored support vectors, which allows the new coming sample to push the decision function away from it, while the compensation of the coefficients of previously stored support vectors with oppositive label to the new sample allows them to push the decision function back to the new sample.

Update Buffer. When the buffer is not full, the new coming sample will be considered as a new support vector and stored in the buffer directly. When the buffer is full, we need to update the buffer correspondingly, which is a very challenging work. In this paper, we investigate several *stream oblivious* policies [8, 10]:

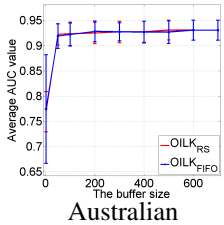
- First-In-First-Out (**FIFO**): With probability 1, we replace the oldest sample in the buffer with $\mathbf{z}_t$. This strategy is simple and intuitive as the coefficient of the oldest sample may be decayed more.
- Reservoir Sampling (**RS**): With probability $\frac{N}{t}$, we update the buffer by randomly replacing one instance in B_t with $\mathbf{z}_t$. This is a widely used strategy in data streaming community and contains the good property that the instances in the buffer simulate a uniform sampling in the original dataset [8].

Table 1: Average AUC performance on seven benchmark datasets

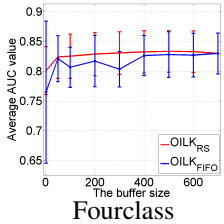
Dataset	Perceptron	OAMseq	OAMgra	NORMA	OILK$_{FIFO}$	OILK$_{RS}$
sonar	0.868 ± 0.056	0.857 ± 0.036	0.856 ± 0.045	0.828 ± 0.038	**0.933 ± 0.040**	0.929 ± 0.039
australian	0.919 ± 0.022	0.925 ± 0.023	0.925 ± 0.021	0.925 ± 0.021	**0.928 ± 0.020**	0.925 ± 0.021
heart	0.898 ± 0.035	**0.912 ± 0.028**	**0.912 ± 0.028**	0.910 ± 0.035	0.907 ± 0.030	0.905 ± 0.028
ionosphere	0.937 ± 0.031	0.927 ± 0.036	0.928 ± 0.036	0.925 ± 0.041	0.948 ± 0.023	**0.954 ± 0.021**
fourclass	0.820 ± 0.038	0.823 ± 0.038	0.823 ± 0.038	0.813 ± 0.035	0.817 ± 0.043	**0.829 ± 0.036**
segment	0.983 ± 0.008	**0.999 ± 0.001**	**0.999 ± 0.000**	0.996 ± 0.002	0.997 ± 0.007	0.997 ± 0.003
satimage	0.635 ± 0.029	0.901 ± 0.016	0.901 ± 0.012	0.879 ± 0.016	**0.905 ± 0.013**	0.896 ± 0.024

Dataset	T	d	$\frac{T^-}{T^+}$
sonar	208	60	1.1443
australian	690	14	1.2476
heart	270	13	1.2500
ionosphere	351	34	1.7857
fourclass	862	2	1.8078
segment	2310	19	6.0000
satimage	6435	36	9.2796

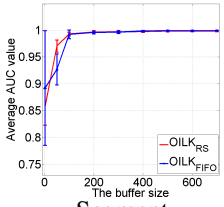
Data description



Australian



Fourclass



Segment

Figure 1: Data description and evaluation of AUC performance with respect to varied buffer sizes.

3 Experimental Results

In this section, we evaluate the empirical performance of the proposed Online Imbalanced Learning with Kernel (OILK) algorithms for imbalanced online learning tasks and compare them with the state-of-the-art online learning algorithms: 1) "Perceptron: the classical Perceptron algorithm [7]; 2-3) "OAM$_{\text{seq}}$ and "OAM$_{\text{gra}}$: the OAM algorithm by the gradient descent and the sequential updating approach [10]; 4) "NORMA": online learning with kernel [5]; 5-6) "OILK$_{\text{FIFO}}$" and "OILK$_{\text{RS}}$": our OILK with the FIFO and the reservoir sampling buffer updating strategy.

Experimental testbed and setup. We conduct experiments on seven benchmark imbalanced datasets randomly selected from machine learning repositories. Due to space limitation, we show the detailed description of the datasets in the table of Fig. 1. We follow the experimental setup in [10] and average AUC results over 20 runs. To make fair comparisons, all algorithms adopt the same setup. For OAM, NORMA, and our OILK, the size of buffer is set to 200. For the OAM algorithm, we apply a 5-fold cross-validation to the training set to find the best penalty parameter $C \in 2^{[-15:1:10]}$. For the NORMA algorithm, we use the default value for the parameters as the authors recommended. Similarly for OILK, we apply a 5-fold cross-validation to select best penalty parameter $C \in 2^{[-15:1:10]}$, best $\eta \in 2^{[-15:1:-5]}$ and $\sigma \in 2^{[-5:1:5]}$ for the width of gaussian kernel.

Results and analysis. From Table 1, we can observe that our proposed OILK attains the best performance in five of the seven datasets. Especially, the results of sonar and satimage outperform other methods significantly. We conjecture when the dataset is too complicated to be classified, our OILK can demonstrate its advantages. Moreover, our OILK beats NORMA nearly all the cases, which implies that AUC maximization plays its effect on imbalanced data learning. In Fig. 1, we also show three typical results of AUC score of our OILK with respect to varied buffer sizes. The results show that the performance increases gradually with the increase of the buffer size in the beginning and does not improve when the size is relative large. This implies that there may exist an optimal value for the buffer size.

4 Conclusion

In this paper, we study the streaming imbalanced learning problem and propose a online imbalanced learning with kernels model to exploit the non-linearity and heterogeneity of the imbalanced data. We optimize the AUC score via the kernel trick and investigate two stream oblivious buffer updating policies to resolve the computational and storing burden. We have shown the properties of our proposed OILK through systematical empirical evaluation. Our work also inspires us further investigation on deriving the generalization bound and the buffer updating strategies.

Acknowledgments

The work described in this paper was fully supported by the Basic Research Program of Shenzhen (Project No. JCYJ20120619152419087 and JC201104220300A), and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 413212 and CUHK 415212).

References

[1] Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. In *NIPS*. MIT Press, 2003.

[2] Wei Gao, Rong Jin, Shenghuo Zhu, and Zhi-Hua Zhou. One-pass auc optimization. In *ICML*, pages 906–914, 2013.

[3] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic(ROC) curve. *Radiology*, 143(1):29–36, 1982.

[4] Purushottam Kar, Bharath K. Sriperumbudur, Prateek Jain, and Harish Karnick. On the generalization ability of online learning algorithms for pairwise loss functions. *CoRR*, abs/1305.2505, 2013.

[5] Jyrki Kivinen, Alexander J. Smola, and Robert C. Williamson. Online learning with kernels. *IEEE Transactions on Signal Processing*, 52(8):2165–2176, 2004.

[6] Francesco Orabona, Joseph Keshet, and Barbara Caputo. The projectron: a bounded kernel-based perceptron. In *ICML*, pages 720–727, 2008.

[7] F. Rosenblatt. The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958.

[8] Jeffrey S. Vitter. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, March 1985.

[9] Yuyang Wang, Roni Khardon, Dmitry Pechyony, and Rosie Jones. Generalization bounds for online learning algorithms with pairwise loss functions. *Journal of Machine Learning Research-Proceedings Track*, 23:13–1, 2012.

[10] Peilin Zhao, Steven C. H. Hoi, Rong Jin, and Tianbao Yang. Online auc maximization. In *ICML*, pages 233–240, 2011.