

An Empirical Study on Budget-Aware Online Kernel Algorithms for Streams of Graphs

Giovanni Da San Martino^a, Nicolò Navarin^{b,*}, Alessandro Sperduti^b

^a*Qatar Computing Research Institute, HBKU, P.O. Box 5825 Doha, Qatar*

^b*Department of Mathematics, University of Padova, via trieste 63, Padova, Italy*

Abstract

Kernel methods are considered an effective technique for on-line learning. Many approaches have been developed for compactly representing the dual solution of a kernel method when the problem imposes memory constraints. However, in literature no work is specifically tailored to streams of graphs. Motivated by the fact that the size of the feature space representation of many state-of-the-art graph kernels is relatively small and thus it is explicitly computable, we study whether executing kernel algorithms in the feature space can be more effective than the classical dual approach. We study three different algorithms and various strategies for managing the budget. Efficiency and efficacy of the proposed approaches are experimentally assessed on relatively large graph streams exhibiting concept drift. It turns out that, when strict memory budget constraints have to be enforced, working in feature space, given the current state of the art on graph kernels, is more than a viable alternative to dual approaches, both in terms of speed and classification performance.

Keywords: online learning, graph kernels, graph streams, online passive aggressive

*Corresponding author

Email addresses: gmartino@qf.org.qa (Giovanni Da San Martino),
mnavarin@math.unipd.it (Nicolò Navarin), sperduti@math.unipd.it (Alessandro Sperduti)

1. Introduction

The amount of data generated in different areas by computer systems is growing at an extraordinary pace, mainly due to the advent of technologies related to the web, ubiquitous services and embedded systems that aim at monitoring the environment in which they are immersed in. Data are, in some cases, generated at a constant rate by sources that can potentially emit an unbounded sequence of elements, i.e. data streams. The processing of data streams requires special care from a computational point of view, since only bounded time and memory resources might be available. Indeed, online algorithms may be required to scale linearly with the number of data items and use a constant, a priori determined, amount of memory (budget). An example of a learning task on streams is binary classification, where the goal is to approximate a function $f : \mathbb{X} \rightarrow \{-1, 1\}$ which partitions the input domain \mathbb{X} into two classes. When dealing with streams, it was early recognized that they tend to evolve with time, giving rise to the well known *concept drift* phenomenon [1], which consists in the function $f()$ changing over time.

In this paper, we focus on graph streams, which involve a large range of application tasks such as chemical compound or image classification (see Sections 4.1.1 and 4.1.2, respectively), as well as malware detection [2], where executables codes represent graph nodes and control flow instructions and API calls represent edges, and Fault Diagnosis in Sensor Networks [3]. Note that we assume that the source generating the stream emits one graph at a time (i.e., we do not have an edge stream as, for example, in [4]).

The traditional approach when dealing with structured data is to transform the data into a suitable vectorial representation. When the examples are graphs, the mapping is commonly referred to as graph embedding [5]. The drawbacks of this approach are that the embedding is task-dependent, and generally computationally expensive. Moreover, the dimensionality of the vector in which the mapping is performed has to be fixed a-priori (see e.g. [6]), and it is the same for all examples ignoring the differences in the intrinsic complexity of each graph.

A viable alternative to graph embedding is the application of graph kernel methods, which is the approach we consider in this paper. Kernel methods are considered state of the art techniques for classification tasks [7, 8, 9, 10]. The class of kernel methods comprises all those learning algorithms that do not require an explicit representation of the inputs but only information

about the similarity between them. The primal version of a kernel method maps the data onto a vectorial feature space (possibly infinite-dimensional): the similarity can be expressed as a dot product in such space. Any kernel method has a correspondent dual version in which each dot product in feature space is replaced by the evaluation of a correspondent kernel function defined on the input space. The great advantage of kernel methods is the fact that the space and time complexity depends on the kernel function and not on the size of the corresponding feature space. Consequently, the size of the model, i.e. the space needed by the learning algorithm for representing its current solution, is defined in terms of a subset of input examples instead of a subset of features. It is recognized that, when the model is expressed as a set of examples, its size tends to grow proportionally to the number of instances emitted by the stream [11]. Various approaches have been defined to limit the size of the model [12, 13, 14]. However, their application to graph data has been practically limited due to the fact that kernels for graphs tend to be computationally very expensive [15, 16, 17]. Recently a few kernels for graphs have been defined which are both efficient and have very competitive performances on many benchmark datasets [18, 19, 10]. Their complexity ranges from linear in the number of edges [18] to a logarithmic factor above linear in the number of nodes [10], thus they might be ideal candidates for being employed on data streams. One of their key characteristics is that they lead to models that can be represented compactly in the primal space. Thus, for these kernels, both techniques defined for the primal and dual space can be effectively exploited.

The main goal of the paper is to study which of the two approaches is best suited for graph streams. We empirically study the behavior of three different algorithms defined in the primal or in the dual space, using the state-of-the-art graph kernels described in [18, 19, 10] and with multiple techniques for managing the budget. We show experimental results on reasonably large real-world datasets and in the presence of a (controlled) concept drift. The results suggest that, under specific budget constraints, working in the primal space is faster and leads to better or comparable results with respect to the classic dual approach.

The paper is organized as described in the following. Section 2, after introducing some notation, recalls important background notions for understanding the paper: graph kernels, online learning algorithms on a budget defined in primal or dual space. Section 3 extends the previously presented online learning algorithms to graph data and discusses several model-pruning

strategies to ensure that strict budget constraints are satisfied. Section 4 studies the performances of the learning algorithms on a budget with respect to the various model strategies and kernel functions. Finally, Section 5 draws conclusions.

2. Background

This section introduces the concepts and algorithms used in the remainder of the paper. We start by introducing some notation in Section 2.1. Section 2.2 briefly reviews kernel functions for graphs outlining the fact that some of the state-of-the-art ones have both low computational complexity and a compact representation as a set of features. Motivated by this last observation, we describe state-of-the-art kernel methods for online learning and budget management techniques working in the dual space, in Section 2.3, and online learning algorithms working directly in feature space, in Section 2.4.

2.1. Notation

A graph $G(V, E, L)$ is a triplet where V is the set of vertices, E the set of edges and $L()$ a function mapping nodes to a set of labels A . A proper subgraph $G_2 = (V_2, E_2, L)$ of $G_1 = (V_1, E_1, L)$ is a graph for which $V_2 \subseteq V_1$, $E_2 = E_1 \cap (V_2 \times V_2)$. A directed acyclic graph (DAG) is a graph where edges are directed and no directed cycle is present. A *proper rooted substructure* of a DAG D is defined in this paper as a subgraph of D obtained by considering a node v of D and all the nodes which can be reached from v using the directed edges of D . A tree is a directed acyclic graph where each node has at most one incoming edge. A proper subtree rooted at node v comprises v and all its descendants. We denote with ρ the maximum outdegree of a graph.

2.2. Graph Kernels

In order to apply a kernel method to graph data, an appropriate kernel function must be provided. Such function, defined on any pair of instances of a domain must be symmetric positive semidefinite. Various similarity measures can be exploited to define a kernel for graphs. For example, a similarity score can be given by the number of subgraphs that two graphs G_1 and G_2 share. Unfortunately, the implementation of this simple idea is very expensive from a computational point of view since recognizing if a subgraph g_1 of G_1 is isomorphic to a subgraph g_2 of G_2 requires to solve a *subgraph isomorphism problem*, which is known to be NP-Complete [15].

Most of the research on graph kernels proceeded by focusing on a restricted class of substructures for which the membership to a graph can be decided in polynomial time (e.g., walks [15, 20, 21], shortest paths [16, 22], subtree patterns [17], small-sized subgraphs [23]) with the aim of obtaining a feature space as large as possible. However, the complexity of the cited algorithms spans from $O(n^3)$ to $O(n^6)$ ¹, where n is the size (number of nodes) of the graphs, which make them hardly applicable to on-line learning tasks with strict time constraints.

Recently, a few kernels with complexity $O(m)$, where m is the number of edges, or $O(n \log n)$, have been defined [18, 19, 10]. Despite their low complexity their performance is considered state of the art on many benchmark datasets. Moreover, their low complexity allow them to be applied to very large datasets. The *Weisfeiler-Lehman subtree kernel* [18] considers the number of subtree patterns (subtrees where every node in the original graph may appear multiple times) up to a fixed height h . This kernel can be computed in $O(hm)$ time on a pair of graphs G_1 and G_2 , where $m = \max(|E_1|, |E_2|)$. Note that the h is a kernel parameter and the authors always use a constant value, so the complexity practically is $O(m)$. The *Neighborhood subgraph pairwise distance kernel* (NSPDK) [19] decomposes a graph into pairs of small subgraphs of radius at most h , up to a maximum distance d : every feature in the explicit feature space represents two particular subgraphs being at a certain distance. Here d and h are kernel parameters which, in order to reduce the computational burden of the kernel evaluation, in practice are kept constant [19]. Finally, the ODD_{ST} kernel, a member of the Ordered Decompositional DAGs Kernel family for graphs [10], decomposes a graph of n nodes into n DAGs. Each DAG is obtained performing a breadth first visit of the graph, up to a fixed height h set by the user, and removing the nodes inducing a cycle. The features associated with a graph are the proper rooted substructures of each DAG.

The set of non-zero features related to the *Weisfeiler-Lehman subtree*, the *Neighborhood subgraph pairwise distance* and the ODD_{ST} kernels, and consequently the associated models, tend to have a compact representation. The number of features generated for a graph is *at most*: nh for the *Weisfeiler-*

¹The kernel in [23] can be computed in $O(n\rho^{k-1})$, where k is the size of the considered subgraphs, on unlabeled graphs. However, in this paper we deal with labeled graphs and the complexity of the kernel for this case is $O(n^k)$.

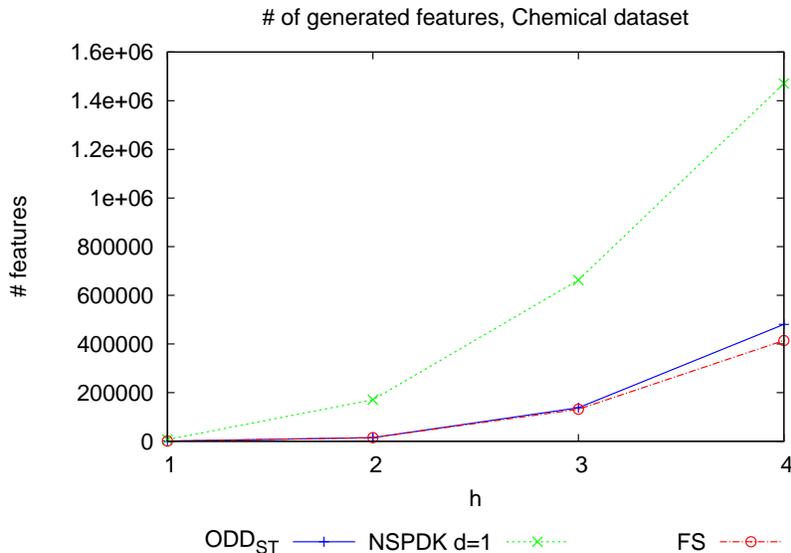


Figure 1: Cumulative number of (different) features generated over the *Chemical* stream according to the ODD_{ST} , NSPDK and FS kernels, for different h parameter values.

Lehman subtree kernel [18]; $\frac{hn\rho^d}{2}$ for NSPDK, where $\frac{n\rho^d}{2}$ is an upper bound on the number of pairs of nodes that are at most at distance d ; $n\rho^h$ for ODD_{ST} [10].

Note that the kernel parameters h , d are assumed to be constant [18, 19, 10] and that, in many practical applications, ρ can be considered constant as well, thus the number of features generated by the different kernels is practically linear. This property will be exploited by the online learning algorithms described in Section 2.4.

Nonetheless, if we consider the size of the feature space induced by the kernels on a whole dataset, the number of different features that are generated may be very high. Figure 1 shows the size of the induced feature space for one of the datasets we will adopt in the experimental part of the paper (see Section 4.3), for different values of the h parameter, for the considered kernels.

2.3. Dual Online Kernel Methods On a Budget

The majority of online kernel methods on a budget are a variant of the perceptron [24] and thus share a common structure. Let us assume the input stream is formed by pairs $e_t = (x_t, y_t)$, where $x_t \in \mathbb{X}$ is an input instance and

$y_t = \{-1, 1\}$ is its label². The goal is to find a hypothesis $h : \mathbb{X} \rightarrow \{-1, 1\}$ such that the expected value of the adopted error measure on the stream is minimized. In the version of the perceptron we introduce here, which we call *Dual* since it is expressed in the kernel dual space (input space), the hypothesis is represented by a subset M of the input instances [12]. M is commonly referred to as the model. The following is a general scheme of the *Dual* version of the perceptron:

Algorithm 1 A general *Dual* perceptron-style algorithm for online kernel learning on a budget.

```

1: Input:  $\beta$  (algorithm dependent),  $B$  (budget size)
2: Initialize  $M$ :  $M = \{\}$ 
3: for each round  $t$  do
4:   Receive an instance  $x_t$  from the stream
5:   Compute the score of  $x_t$ :  $S(x_t) = \sum_{i=1}^{|M|} y_i \tau_i K(x_i, x_t)$ 
6:   Receive the correct classification of  $x_t$ :  $y_t$ 
7:   if  $y_t S(x_t) \leq \beta$  ( $x_t$  incorrectly classified) then
8:     while  $|M| + |x_t| > B$  do
9:       select an element  $x_j \in M$  for removal
10:       $M = M \setminus \{x_j\}$ 
11:     end while
12:     update the hypothesis:  $M = M \cup \{(y_t \tau_t, x_t)\}$ 
13:   end if
14: end for

```

In Algorithm 1, $|M|$ represents the size of the model, i.e. the sum of the size of the instances in M . In the same way $|x_t|$ is the size of x_t . If the input instances are vectorial data, their size is constant, thus in order to add an element to M , it is sufficient to remove only one instance from M , i.e. the while loop in Algorithm 1-line 8 is executed exactly once. As it will be detailed in Section 3, this is not the case in our scenario where the input instances are graphs and their size is not constant. Note that Algorithm 1 tries to use as much memory as it is allowed to (without exceeding the limit B): line 8 shows that one example would be removed from the model only if the algorithm, by inserting a novel example in the model, exceeded the memory limit B . In all other cases, any new erroneously-classified example is inserted in the model (line 12). All we shall see, the same observation will apply to the two other algorithms presented in this paper.

Many online algorithms can be seen as instances of Algorithm 1. For ex-

²As in the standard online setting, we assume that the target value y_t is observed only after the system has predicted an output for x_t .

ample, by setting $B = \infty$, $\tau = 1$, $\beta = 1$, we obtain the dual perceptron [12]. The Online Passive-Aggressive algorithm [13] tries to select an hypothesis with a unit margin on the examples. It is obtained with $B = \infty$, $\beta = 1$, $\tau_i = \min \left\{ C, \frac{1-S(x_i)}{K(x_i, x_i)} \right\}$, where C is a user-defined non-negative parameter. In [14] it is described an update rule which tries to project the new instance onto the span of the current support set M . The resulting hypothesis is compared to the one obtained by inserting the whole instance into the model: if the difference between the two hypotheses is not greater than a user-defined threshold, then only the projected instance is added to the model. Computing the projection requires quadratic time and space with respect to the size of the support set, thus severely limiting the application to graph streams. Since the three algorithms assume $B = \infty$, no elements are removed from M . Thus, even if they try to minimize the size of the model, they do not provide any strategy to ensure that such size will not exceed any a priori given budget.

When the problem setting imposes a budget B on the size of the model, various strategies can be employed for selecting which elements should be removed from M . In [25] the elements to be removed are chosen randomly. The Forgetron removes the oldest example in M [26]: a decay factor is applied to the τ values in such a way that the oldest examples in M have lower and lower impact on the computation of $S()$. Crammer et al. [27] proposed to remove from M any redundant example, i.e. the example with least impact on the margin of the hypothesis. This approach, however, is computationally expensive and thus it is not suitable for processing high dimensional data streams. In [28] the Online Passive-Aggressive algorithm [13] has been extended to handle budget constraints. The idea is to modify the update rule such that the resulting hypothesis, after decreasing the model size such that the budget constraint is respected, has a small loss on the new example and it is similar to the current hypothesis. They describe three algorithms of increasing complexity and efficacy: BPA-S, BPA-NN, BPA-P. Among these, BPA-S has linear space and time complexity with respect to the model size.

2.4. Primal Algorithms for Online Learning On a Budget

By the properties of kernel functions, each kernel evaluation corresponds to a dot product in an associated feature space. Then Algorithm 1 has a corresponding version in feature space in which the examples are represented by their projection in feature space $\phi(x_t) \in \mathbb{R}^s$ (with s being the size of the

feature space). The hypothesis is represented by a vector $w \in \mathbb{R}^s$ [24], where the elements of M are replaced by their sum:

$$w = \sum_{\phi(G_j) \in M} y_j \tau_j \phi(G_j). \quad (1)$$

The score is computed as $S(x_t) = w_t \cdot \phi(x_t)$ and the hypothesis is updated as $w_{t+1} = w_t + \tau_t y_t \phi(x_t)$. Given the fixed size of w , the standard perceptron does not take into account budget constraints. We refer to such version as *Primal*.

An algorithm, similar to the one just described, has been presented in [29]: the update step is a stochastic gradient descent rule followed by a rounding step in which the small coefficients are set to zero. Since zero features may not be explicitly represented, the rounding phase allows to reduce the model size. In [30] a framework for minimizing a convex loss function together with a convex regularization term is presented. The update rule is constituted by two phases: the first one is a subgradient step with respect to the loss function and the second one looks for a vector which maximizes the similarity to the one obtained in the first phase while minimizing the regularization term. Various instantiations are discussed: among these, the one making use of the ℓ_1 norm as a regularization term is interesting for this paper, since it promotes sparse solutions. Note that the literature on online learning algorithms working directly in feature space is incredibly vast, but here we are interested in algorithms corresponding to state of the art dual approaches. Indeed, our purpose is to assess the viability of primal approaches in the context of kernel methods.

As for the algorithms discussed in Section 2.3, a drawback of the algorithms listed in this section is that, they do not provide any strategy to ensure that the size of the model w will not exceed any a priori given budget.

3. Budget-aware Algorithms for Structured Data

In this paper, we study three algorithms, together with different strategies for managing the budget, for graph streams. Our first proposal, Algorithm 1, needs a few adaptations before it can be applied to graph data. Given the variable size of graph data we make use of the following measure for computing the size of the model in Algorithm 1:

$$|M| = \sum_{G_j \in M} (|V_{G_j}| + |E_{G_j}| + 1), \quad (2)$$

where the constant term 1 takes into account the occupancy of the value $\tau_t y_t$. The removal rule in Algorithm 1 is modified as follows: when G_t has to be inserted, instances are removed from M until $|M| + |V_{G_t}| + |E_{G_t}| + 1 < B$, where $|M|$ is computed according to eq. (2).

The time complexity of an online algorithm depends on the number of graphs in M and the complexity of the kernel function employed. In those settings in which the number of features associated with a kernel is not significantly greater than the size of the input, the evaluation of the kernel function may be greatly speeded up if it is performed as dot product of the corresponding feature vectors. Examples of kernels having such property are [18, 19, 10]. In the remainder of the section our observations will be restricted to this class of kernels. The actual size of vectors $\phi(G)$ can be much less than s if only non-null elements of $\phi(G)$ are represented in sparse format. We will refer to the number of non-null features of $\phi(G)$ as $|\phi(G)|$. These observations lead to the *Primal/Dual* algorithm (referred to as *mixed* in the following):

Algorithm 2 *Mixed* perceptron-style algorithm for online learning on a budget.

```

1: Input:  $\beta$  (algorithm dependent),  $B$  (budget size)
2: Initialize  $M$ :  $M = \{\}$ 
3: for each round  $t$  do
4:   Receive an instance  $G_t$  from the stream
5:   Compute the score of  $G_t$ :  $S(G_t) = \sum_{\phi(G_j) \in M} y_j \tau_j \phi(G_j) \cdot \phi(G_t)$ 
6:   Receive the correct classification of  $G_t$ :  $y_t$ 
7:   if  $y_t S(G_t) \leq \beta$  ( $G_t$  incorrectly classified) then
8:     update the hypothesis:
9:     while  $1 + \sigma |\phi(G_t)| + \sum_{\phi(G_j) \in M} 1 + \sigma |\phi(G_j)| > B$  do
10:       select an element  $\phi(G_j) \in M$  and remove it:  $M = M \setminus \{\phi(G_j)\}$ 
11:     end while
12:      $M = M \cup \{y_t \tau_t \phi(G_t)\}$ 
13:   end if
14: end for

```

Note that the model size is computed as $\sum_{\phi(G_j) \in M} 1 + \sigma |\phi(G_j)|$, where the constant 1 accounts for the $y_t \tau_t$ value and σ is the memory occupancy of a feature: if $\phi(G)$ is represented in sparse format as pairs $(i, \phi_i(G))$, where $\phi_i(G)$ is the value of the i -th feature of G , then $\sigma = 2$. As we will see in Section 3.1, while σ might be influenced by the budget management strategy employed, in all the experiments performed in this paper with Algorithm 2 the value σ will remain unchanged.

Since in Algorithm 2 the projection $\phi(G)$ is not computed for every kernel

evaluation, Algorithm 2 is expected to be faster than Algorithm 1. However, if $|\phi(G)| > |V_{G_j}| + |E_{G_j}|$, which generally holds, it uses more memory.

Finally, we introduce a budget online algorithm working in feature space. The idea is to replace all elements of M with their sum as in eq. (1). However, by so doing, we lose the connection between features and the instances they belong to. As a consequence, during the update of the hypothesis it is no more possible to select a whole vector $\phi(G)$ for removal. Thus we propose to remove single features from w when $|w| > B$ (here $|w|$ is the total number of non-null features appearing in any example added to the model).

Algorithm 3 *Primal* perceptron-style online learning on a budget.

```

1: Input:  $\beta$  (algorithm dependent)
2: Initialize  $w$ :  $w_0 = (0, \dots, 0)$ 
3: for each round  $t$  do
4:   Receive an instance  $G_t$  from the stream
5:   Compute the score of  $G_t$ :  $S(G_t) = w_t \cdot \phi(G_t)$ 
6:   Receive the correct classification of  $G_t$ :  $y_t$ 
7:   if  $y_t S(G_t) \leq \beta$  ( $G_t$  incorrectly classified) then
8:     while  $\sigma|w + \phi(G_t)| > B$  do
9:       select a feature  $i$  and remove it from  $w$ 
10:    end while
11:   update the hypothesis:  $w_{t+1} = w_t + \tau_t y_t \phi(G_t)$ 
12:   end if
13: end for

```

The total memory occupancy of the model in Algorithm 3 reduces to $\sigma|w|$.

Note that the elimination of the set M allows Algorithm 3 to save a significant amount of memory while still being faster than Algorithms 1 and 2.

3.1. Budget Management

We have left unspecified how to select the examples/features to be removed when the budget is full in Algorithms 1-3. As we briefly discussed in Section 2.3, complex strategies, which would require to solve an optimization problem, are usually expensive from the computational point of view [27, 28]. This is especially true for the graph domain for two main reasons. Graph data are generally high-dimensional thus making the solution of the optimization problems even more computationally expensive. The second reason is that, for instance the problem solved in [28] (eq. 7) assumes that removing one example frees enough space for the novel example to be inserted, which does not hold for graphs since they are of variable size. Modifying

the optimization problem to account for the removal of a subset of examples would increase the complexity of the problem, and the resulting method would not respect the constraint of linear processing time imposed by the setting considered in this paper. For such reasons, we focus in this paper on heuristics for selecting the elements to be removed from the model. Given the differences in how the model is represented in the three algorithms, different strategies for pruning the model can be applied. We have explored the following strategies for Algorithms 1 and 2:

- “random”, examples are removed randomly with uniform probability;
- “oldest”, the oldest examples are removed;
- “ τ ”, the examples with lowest τ values are removed. If more than one example has such τ value, the candidate is randomly selected.

Note that the implementation of the three strategies does not increase the memory occupancy of the model.

Since any kernel method using the kernel functions in [18, 19, 10] can be performed in the primal space, it is possible to apply *feature selection* techniques, i.e. deleting non-informative features, in order to reduce noise in the data and the size of the model [31]. A typical approach is to compute a statistical measure for estimating the relevance of each feature with respect to the target concept, and to discard the less-correlated features. Before describing the strategies for pruning the model for Algorithm 3, we introduce an example of such measure, the F-score [31]. In the traditional batch scenario, the F-score of a feature i is defined for binary classification tasks as follows:

$$Fs(i) = \frac{(AVG_i^+ - AVG_i)^2 + (AVG_i^- - AVG_i)^2}{\frac{\sum_{j \in Tr^+} (f_i^j - AVG_i^+)^2}{|Tr^+| - 1} + \frac{\sum_{j \in Tr^-} (f_i^j - AVG_i^-)^2}{|Tr^-| - 1}} \quad (3)$$

where AVG_i is the average value of feature i in the dataset, AVG_i^+ (AVG_i^-) is the average value of feature i in positive (negative) examples, $|Tr^+|$ ($|Tr^-|$) is the number of positive (negative) examples and f_i^j is the value of feature i in the j^{th} example of the dataset. Features that get small values of F-score are

not very informative with respect to the binary classification task ³. Eq. (3) cannot be applied as is to a stream since instances arrive one at time. As a minor contribution, we rewrite an incremental version of the F-score. Let \mathcal{I}_t^+ (\mathcal{I}_t^-) be the set of positive (negative) instances which have been observed from the stream after having read t instances, then the F-score $Fs(i, t)$ can be rewritten by using the following quantities:

$$\begin{aligned} n_t^+ &= |\mathcal{I}_t^+|, & f_i^+(t) &= \sum_{j \in \mathcal{I}_t^+} f_i^j, & f_i^{2,+}(t) &= \sum_{j \in \mathcal{I}_t^+} (f_i^j)^2 \\ n_t^- &= |\mathcal{I}_t^-|, & f_i^-(t) &= \sum_{j \in \mathcal{I}_t^-} f_i^j, & f_i^{2,-}(t) &= \sum_{j \in \mathcal{I}_t^-} (f_i^j)^2. \end{aligned}$$

In fact, we have:

$$\begin{aligned} AVG_{i,t}^+ &= \frac{f_i^+(t)}{n_t^+}, & AVG_{i,t}^- &= \frac{f_i^-(t)}{n_t^-} \\ AVG_{i,t} &= \frac{f_i^+(t) + f_i^-(t)}{n_t^+ + n_t^-} \end{aligned}$$

and

$$Fs(i, t) = \frac{(AVG_{i,t}^+ - AVG_{i,t})^2 + (AVG_{i,t}^- - AVG_{i,t})^2}{D_t^+ + D_t^-} \quad (4)$$

where

$$\begin{aligned} D_t^+ &= \frac{f_i^{2,+}(t) - 2AVG_{i,t}^+ f_i^+(t) + n_t^+ (AVG_{i,t}^+)^2}{n_t^+ - 1}, \\ D_t^- &= \frac{f_i^{2,-}(t) - 2AVG_{i,t}^- f_i^-(t) + n_t^- (AVG_{i,t}^-)^2}{n_t^- - 1}. \end{aligned}$$

By defining $\delta^+(t+1) = 1$ if the $(t+1)$ th instance is positive; otherwise $\delta^+(t+1) = 0$, and $\delta^-(t+1) = 1 - \delta^+(t+1)$, the quantities of interest can be updated incrementally as follows:

$$\begin{aligned} n_{t+1}^+ &= n_t^+ + \delta^+(t+1), & f_i^+(t+1) &= f_i^+(t) + \delta^+(t+1) f_i^j, \\ f_i^{2,+}(t+1) &= f_i^{2,+}(t) + (\delta^+(t+1) f_i^j)^2. \end{aligned}$$

³Even though F-score is known not to take into account correlation between features, we select that measure for computational reasons.

$$n_{t+1}^- = n_t^- + \delta^-(t+1), \quad f_i^-(t+1) = f_i^-(t) + \delta^-(t+1)f_i^j,$$

$$f_i^{2,-}(t+1) = f_i^{2,-}(t) + (\delta^-(t+1)f_i^j)^2.$$

In order to incrementally compute the F-score, we need to keep track, for each feature i , of the following quantities: $f_i^+(t)$, $f_i^-(t)$, $f_i^{2,+}(t)$, $f_i^{2,-}(t)$.

We have explored the following strategies for Algorithm 3:

- *random* strategy: features are removed randomly with uniform probability. This strategy does not affect the size of the model, which is thus obtained setting $\sigma = 2$ in Algorithm 3.
- *weight*: first, all the features of the example which are already present in the model, are inserted. This maximizes the information of the algorithm without increasing memory occupation. Next, for each feature left f of the example, the feature of the model with lowest absolute w_i value (the weight associated with feature f_i), is selected. Note that if all the features in the model have their w_i higher than f , then f is not inserted. The size of the model when this strategy is employed is obtained setting $\sigma = 2$ in Algorithm 3.
- *oldest* strategy: similar to the *weight* strategy, but in this case we remove the least recently used feature. We need to associate to each feature the time in which that feature has been last inserted/modified. The size of the model is obtained setting $\sigma = 3$.
- *F-score*: it is similar to the *weight* strategy, the only difference being that the w_i value is replaced by the F-score, computed according to eq. (3). By using the incremental version of the F-score, the correct size of the model is obtained by setting $\sigma = 5$ in Algorithm 3, since we need to keep track of the index i and the four valued necessary to incrementally update the F-score.

Note that the F-score strategy has no correspondence for *Mixed* and *Dual* algorithms. This strategy removes from the model the features with the lowest associated F-score. F-score measures the correlation of a feature with the target (+1 or -1). Indeed, a feature can appear in different examples, some positive and some negative. If there is a strong correlation with either class, the F-score of a feature will be high. On the contrary, *Mixed* and *Dual* algorithms remove whole examples from the budget. Since an example have

a single label associated, that can be +1 or -1, it is not possible to compute correlation measures in this case.

4. Experimental results

In this section, we empirically compare Algorithms 1-3 with state-of-the-art kernel functions for graphs described in Section 2.2 and various budget management strategies on two graph datasets: the first one is composed of chemical compounds and the second one is composed of images. Our purpose in this section is to study the performances, both in terms of prediction accuracy and running times, of the three algorithms as the memory budget varies, and to determine which algorithm is more appropriate for each setting.

We start by describing in Section 4.1 how the datasets were obtained. Then, in Section 4.2, we introduce the experimental setup and the adopted evaluation measure. Finally, the obtained results are illustrated and discussed in Section 4.3.

4.1. Dataset Description

4.1.1. Chemical Dataset

We have created graph streams combining two graph datasets available from the PubChem website (<http://pubchem.ncbi.nlm.nih.gov>). PubChem is a source of chemical structures of small organic molecules and their biological activities. It contains the bioassay records for anti-cancer screen tests with different cancer cell lines. Each dataset belongs to a certain type of cancer screen. For each compound an activity score is reported. The activity score for the selected datasets is based on increasing values of $-\text{LogGI}_{50}$, where GI_{50} is the concentration of the compound required for 50% inhibition of tumor growth. A compound is classified as active (positive class) or inactive (negative class) if the activity score is, respectively, above or below a specified threshold. By varying the threshold we were able to simulate a drift on the target concept. Our dataset is a combination of the "AID: 123" and "AID: 109" datasets from PubChem. In "AID:123", growth inhibition of the MOLT-4 human Leukemia tumor cell line is measured as a screen for anti-cancer activity. The dataset comprises 40,876 compounds, each one represented by a graph, tested at 5 different concentrations. The average number of nodes for each graph in this dataset is 26.8, while the average

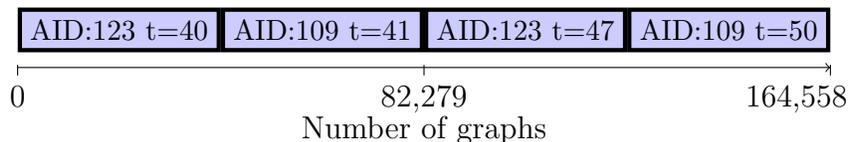


Figure 2: Composition of the stream of graphs on chemical data. Four different target concepts are obtained by using different threshold values (t) on the activity scores of the compounding datasets.

number of edges is 57.68. In “AID:109”, growth inhibition of the OVCAR-8 human Ovarian tumor cell line is measured as a screen for anti-cancer activity on 41,403 compounds. The average number of nodes for each compound is 26.77, while the average number of edges is 57.63. For each dataset, we used two different threshold values to simulate the concept drift: the median of the activity scores and the value such that approximately 3/4 of the compounds are considered dataset to be inactive (negative target). Finally, the stream has been obtained as the concatenation of “AID: 123” with threshold 1, “AID: 109” with threshold 1, “AID: 123” with threshold 2, “AID: 109” with threshold 2 (Figure 2). We call this stream *Chemical*. Note that the stream is composed by four different concepts and comprises a total of 164,558 graphs. Overall, the maximum number of nodes in a graph of the stream is 229, the maximum node outdegree is 6 and the alphabet size is 202. In order to assess the dependency of the results from the order of concatenation of the datasets, we created a second stream as: “AID: 123” with threshold 1, “AID: 123” with threshold 2, “AID: 109” with threshold 1, “AID: 109” with threshold 2. Since the results were very similar to the ones obtained for the first dataset, for the sake of space, we do not report here the results for this second stream. It should be stressed that the selected datasets represent very challenging classification tasks, independently of the value selected as the activity score threshold.

4.1.2. Image Dataset

We created a stream of graphs from the *LabelMe* dataset⁴. The dataset comprises a set of images whose objects are manually annotated via the LabelMe tool [32]. The images are divided into several categories. We have removed those images having less than 3 annotations. We have selected six

⁴<http://labelme.csail.mit.edu/Release3.0/browserTools/php/dataset.php>

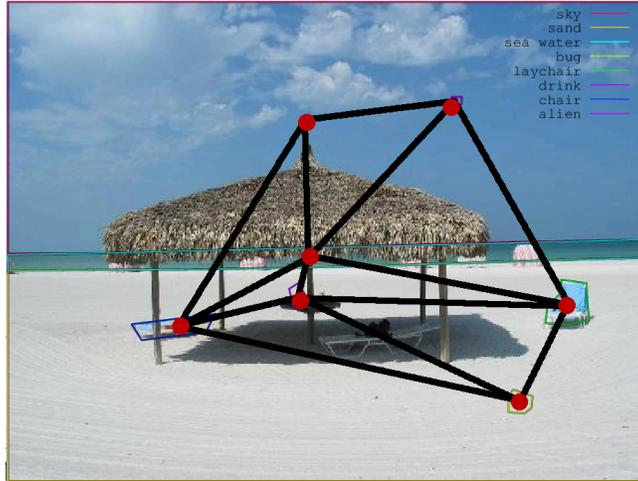


Figure 3: An example of graph construction from an annotated image.

categories amongst the ones having the largest number of images: “office” (816), “home” (928), “houses” (1,294), “urban_city” (865), “street” (1,069), “nature” (370). In total we considered 5,342 images.

We then transformed each image into a graph: the annotated objects of the image become the nodes of the graph. The edges of the graph are determined according to the Delaunay triangulation [33]. The basic idea of the Delaunay triangulation is to connect spatially neighbouring nodes. Figure 3 gives an example of the construction of a graph from an image. The average number of nodes per graph is 14.37 and the average number of edges is 63.61.

The stream is made up of six parts (each part representing a different concept), for each of them one of the categories is selected as the positive class while the remaining ones represent the negative class; in order to simulate concept drifts each one of the 5,342 images appears six times in the stream: once with a positive class label, and 5 times with negative class label. The total number of examples composing the stream is 32,052, the maximum number of nodes of a graph is 201, the maximum node degree is 46 and the alphabet size is 65.

4.2. Experimental setup

For all the considered algorithms (*Primal*, *Mixed* and *Dual*), the β and τ values were instantiated as $\beta = 1$, $\tau_i = \min \left\{ C, \frac{1-S(x_i)}{K(x_i, x_i)} \right\}$, as described in [28] for the BPA-S algorithm. We chose BPA-S among the three BPA algorithms presented in [28], because: *i*) the results in the original paper show that, while being the fastest algorithm, the accuracy with respect to the other BPA versions does not degrade significantly; *ii*) using (also) BPA-P or BPA-NN would have increased significantly the total time required for the experimentation.

The C parameter has been tested in the set $\{0.01, 0.1, 1.0\}$ for both *Chemical* and *Image* datasets. By varying the C value, the results of the comparison between the three algorithms do not change. Therefore we report here only the results related to $C=0.01$. In order to increase the robustness of the results, the three algorithms have been tested with three different graph kernels:

- the *Weisfeiler-Lehman subtree kernel* (FS) [18] with parameter values $h = \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$;
- the *Neighborhood subgraph pairwise distance kernel* (NSPDK) [19] with parameter values $h = \{1, 2, 3, 4\}$, $d = \{1, 2, 3, 4, 5, 6\}$.
- the *ODD_{ST}* kernel [10] with parameter values $\lambda = \{0.8, 1, 1.2, 1.4, 1.6, 1.8\}$, $h = \{1, 2, 3, 4\}$;

All the proposed algorithms have the same upper bound B on memory usage (budget), and the memory occupancy of the algorithms is calculated for *Dual* as in eq. (2), for *Mixed* as of line 9 of Algorithm 2 and for *Primal* as described in line 8 of Algorithm 3 (note that the size of the model for *Primal* also depends on the budget management strategy). We experimented with budget values between 10,000 and 50,000 *memory units* (assuming each memory unit can store a floating point or integer number) for the *Chemical* dataset, and between 1,000 and 100,000 for the *Image* dataset. Higher values, for both datasets, were not tested since the time needed for the *Dual* Algorithm to terminate became excessive (more than 48 hours).

As for the strategies for managing the budget, we focused on the “oldest” and “ τ ” ones for *Dual* and *Mixed* algorithms. We focused on the “oldest” and “weight” strategies for *Primal* algorithm (where we recall that “weight”

is similar in spirit to “ τ ” in the *Primal* setting). Moreover, we considered also the “F-score” strategy for the *Primal* algorithm.

The random strategy has not been implemented because it tends to have worse performances [28].

The class distribution on the streams is unbalanced, therefore the Area Under the Receiver Operating Characteristic (AUROC) and the Balanced Accuracy [31] were adopted as performance measure. The AUROC measure is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, thus it avoids inflated performance estimates on imbalanced datasets. Since the results computed with Balanced Accuracy are very similar to the ones computed with the AUROC, we report only the latter, being the AUROC more popular than the Balanced Accuracy.

The plots in Figures 4-9, Figures 12-17 and Table 1 regarding the AUROC measure are obtained as follows: for each run (Dataset/Kernel/parameters combination) the AUROC measure is sampled every 50 examples. Then we compute the average over all samples and obtain a single value. We chose not to show the behavior of each algorithm during a single run because we have performed more than 300 runs. The running times are computed on a machine with two Intel(R) Xeon(R) CPU E5-4640@ 2.40GHz equipped with 256GB of RAM. Notice that the executions use a single core and a very limited amount of RAM.

4.3. Results and discussion

The aim of the experiments is to compare correspondent budget management strategies for *Primal*, *Dual* and *Mixed*: *i)* *oldest* for the three algorithms; *ii)* *weight* for *Primal* and τ for *Mixed* and *Dual*. For each of the above correspondent budget strategies, we observe the performances of the three algorithms, for each combination of kernel function and kernel parameters, as the budget varies. Section 4.3.1 reports the experiments on the Chemical Dataset. Section 4.3.2 reports the experiments on the Image Dataset. Finally, section 4.3.3 draws general conclusions on the experiments.

4.3.1. Experiments on the Chemical Dataset

The Figures 4-10 report the results for one kernel, one specific budget management policy, two budget values, $B = 10k$ and $B = 50k$. Each Figure is divided into 4 subfigures: the ones on the left side refer to budget $B = 10k$, the ones on the right refer to budget $B = 50k$; the two figures on top report

Oldest policy, Chemical dataset, FS kernel

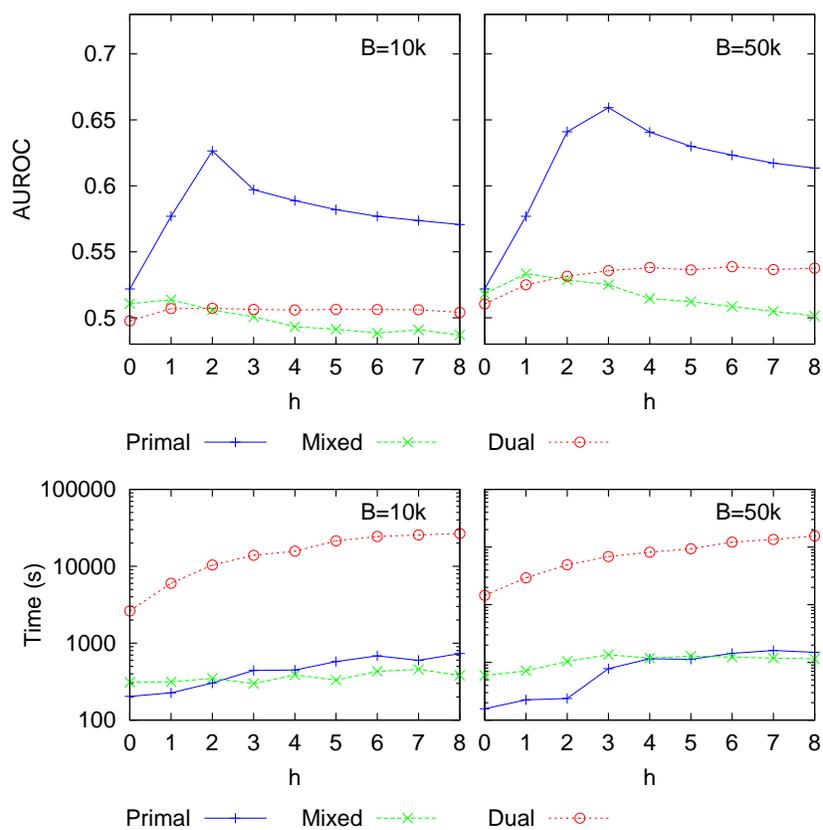


Figure 4: Average AUROC value computed over all stream instances for memory budgets $B = 10k$ (top left) and $B = 50k$ (top right) for algorithms *Primal*, *Mixed* and *Dual* with respect to the values of the FS kernel parameter. Below each of the plots there is a second one with the corresponding running times. The plots refer to the *Chemical* stream and the *oldest* budget maintenance policy.

Weight/ τ policies, Chemical dataset, FS kernel

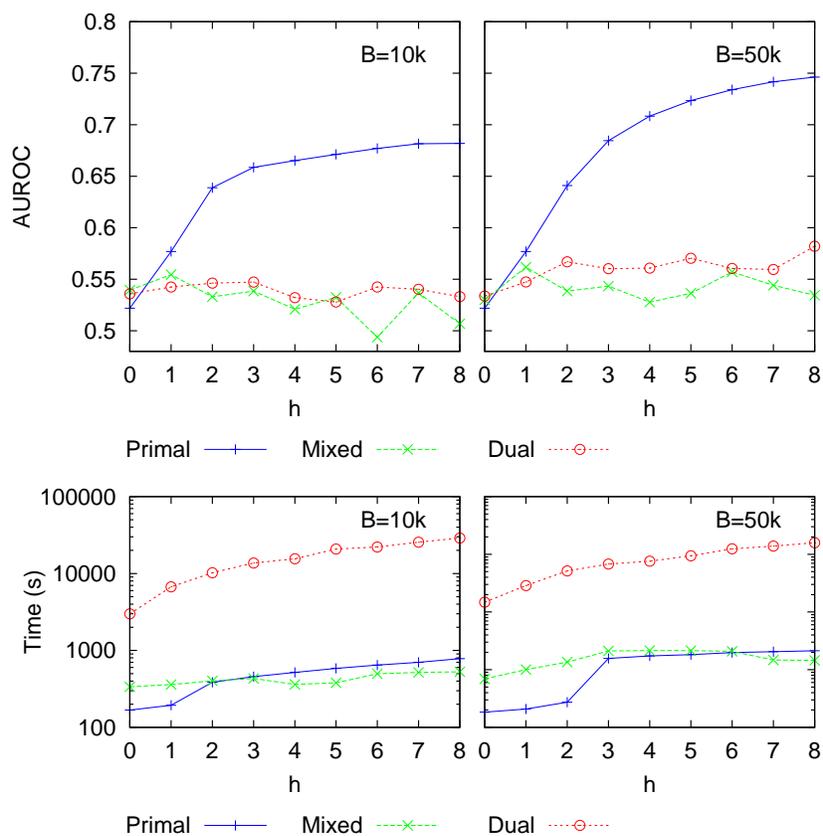


Figure 5: Average AUROC value computed over all stream instances for memory budgets $B = 10k$ (top left) and $B = 50k$ (top right) for algorithms *Primal*, *Mixed* and *Dual* with respect to the values of the FS kernel parameter. Below each of the plots there is a second one with the corresponding running times. The plots refer to the *Chemical* stream and the *weight/ τ* budget maintainance policies.

Oldest policy, Chemical dataset, NSPDK kernel

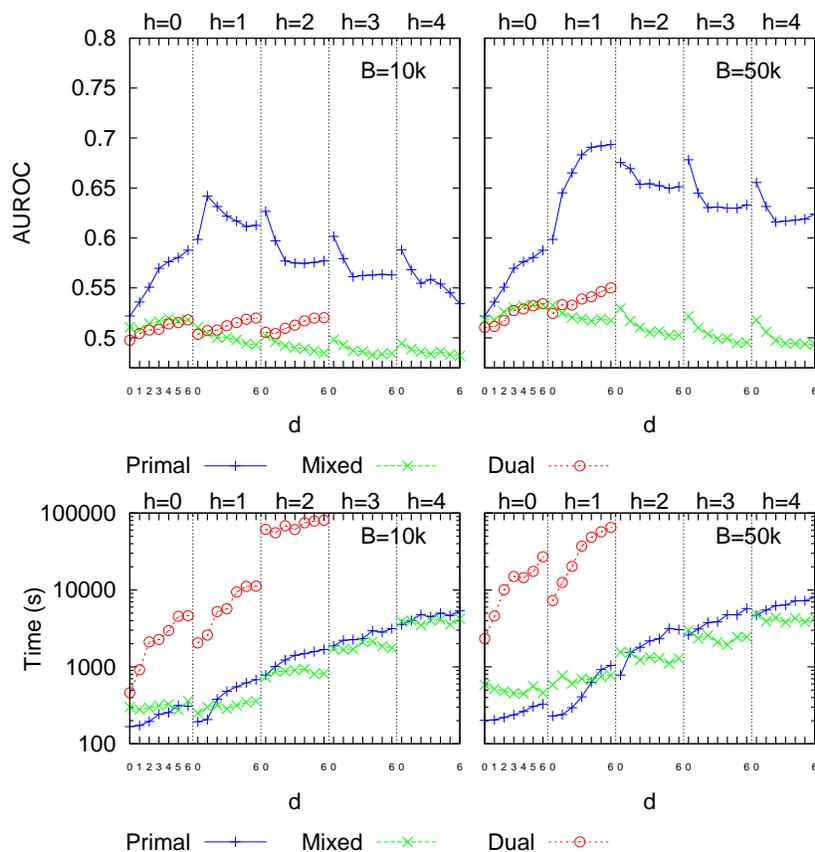


Figure 6: Average AUROC value computed over all stream instances for memory budgets $B = 10k$ (top left) and $B = 50k$ (top right) for algorithms *Primal*, *Mixed* and *Dual* with respect to the values of the NSPDK kernel parameters. Below each of the plots there is a second one with the corresponding running times. The plots refer to the *Chemical* stream and the *oldest* budget maintenance policy. Missing values indicate that the corresponding execution has not terminated in 48 hours.

Weight/ τ policies, Chemical dataset, NSPDK kernel

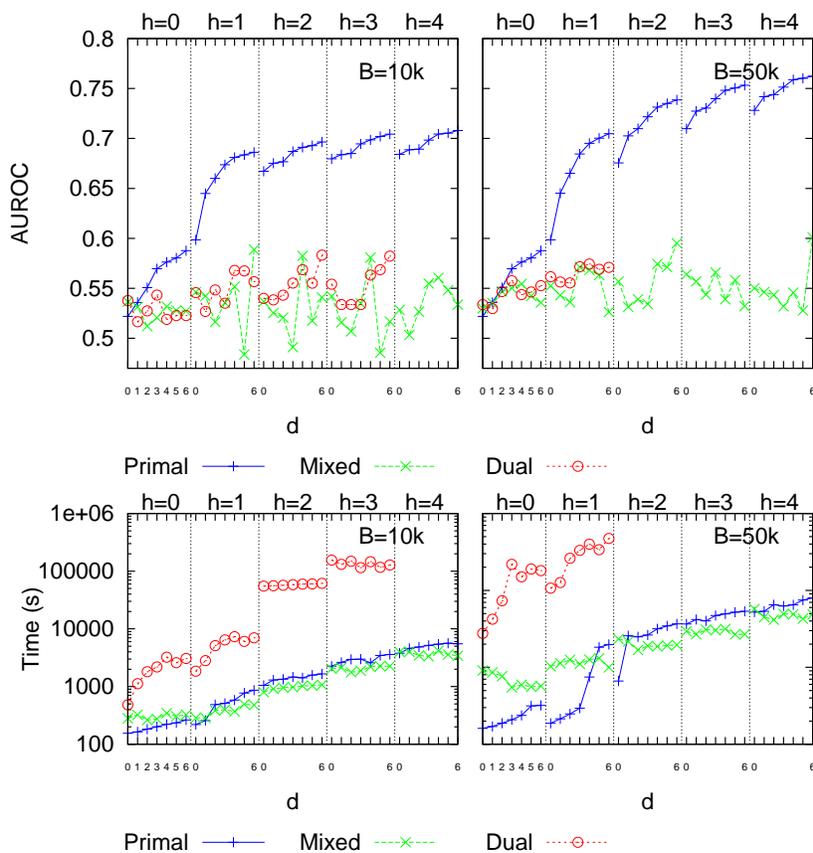


Figure 7: Average AUROC value computed over all stream instances for memory budgets $B = 10k$ (top left) and $B = 50k$ (top right) for algorithms *Primal*, *Mixed* and *Dual* with respect to the values of the NSPDK kernel parameters. Below each of the plots there is a second one with the corresponding running times. The plots refer to the *Chemical* stream and the *weight*/ τ budget maintenance policies. Missing values indicate that the corresponding execution has not terminated in 48 hours.

Oldest policy, Chemical dataset, ODD_{ST} kernel

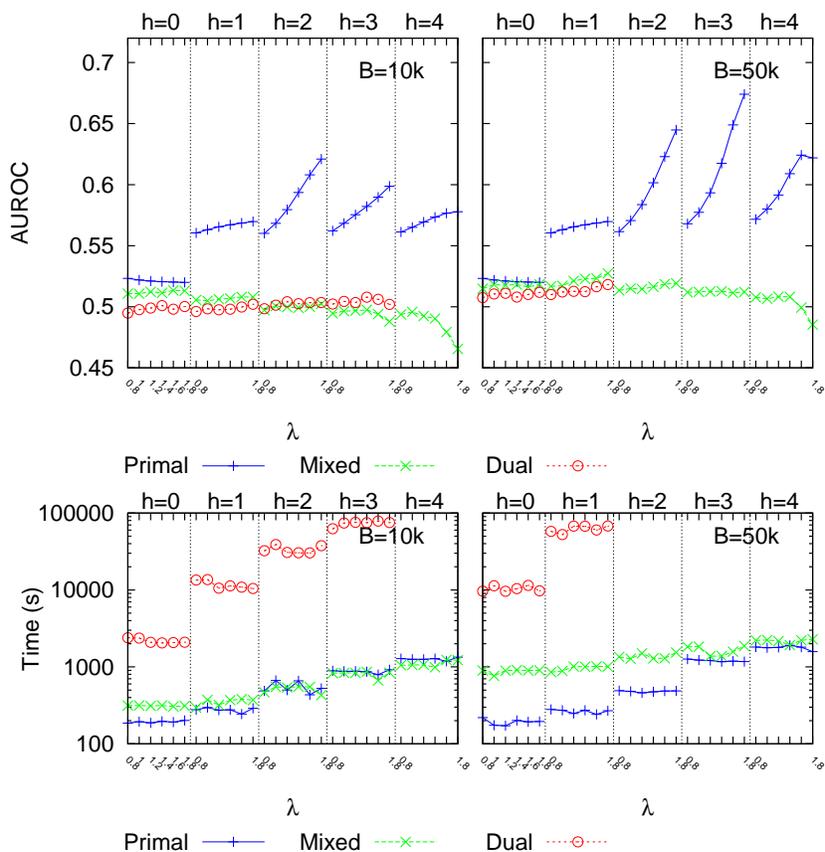


Figure 8: Average AUROC value computed over all stream instances for memory budgets $B = 10k$ (top left) and $B = 50k$ (top right) for algorithms *Primal*, *Mixed* and *Dual* with respect to the values of the ODD_{ST} kernel parameters. Below each of the plots there is a second one with the corresponding running times. The plots refer to the *Chemical* stream and the *oldest* budget maintenance policy. Missing values indicate that the corresponding execution has not terminated in 48 hours.

Weight/ τ policies, Chemical dataset, ODD_{ST} kernel

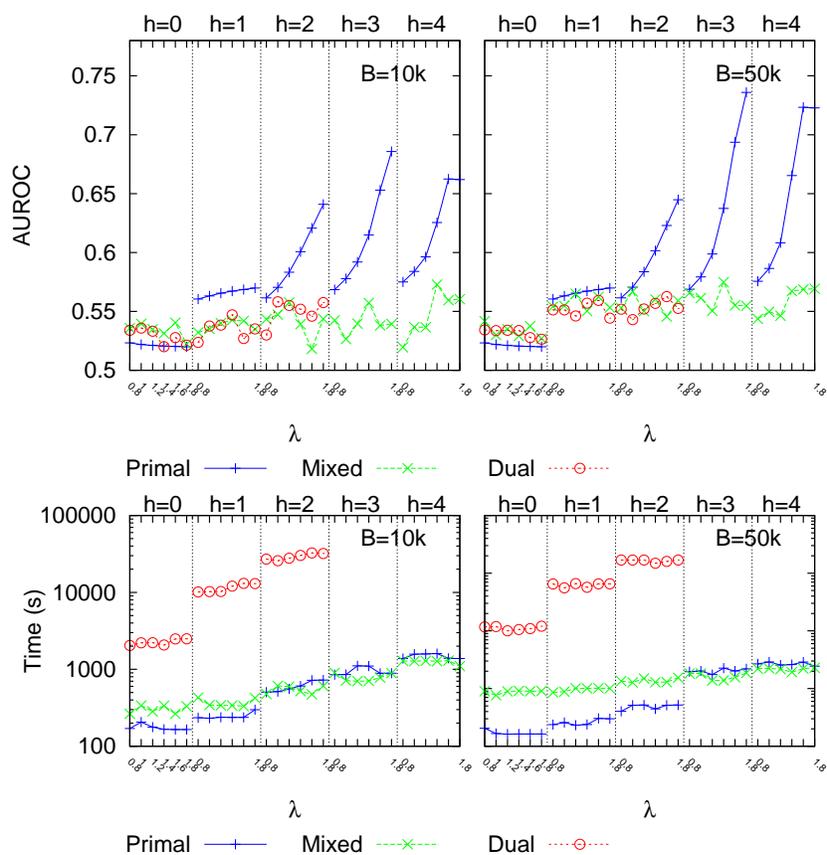


Figure 9: Average AUROC value computed over all stream instances for memory budgets $B = 10k$ (top left) and $B = 50k$ (top right) for algorithms *Primal*, *Mixed* and *Dual* with respect to the values of the ODD_{ST} kernel parameters. Below each of the plots there is a second one with the corresponding running times. The plots refer to the *Chemical* stream and the *weight*/ τ budget maintenance policies. Missing values indicate that the corresponding execution has not terminated in 48 hours.

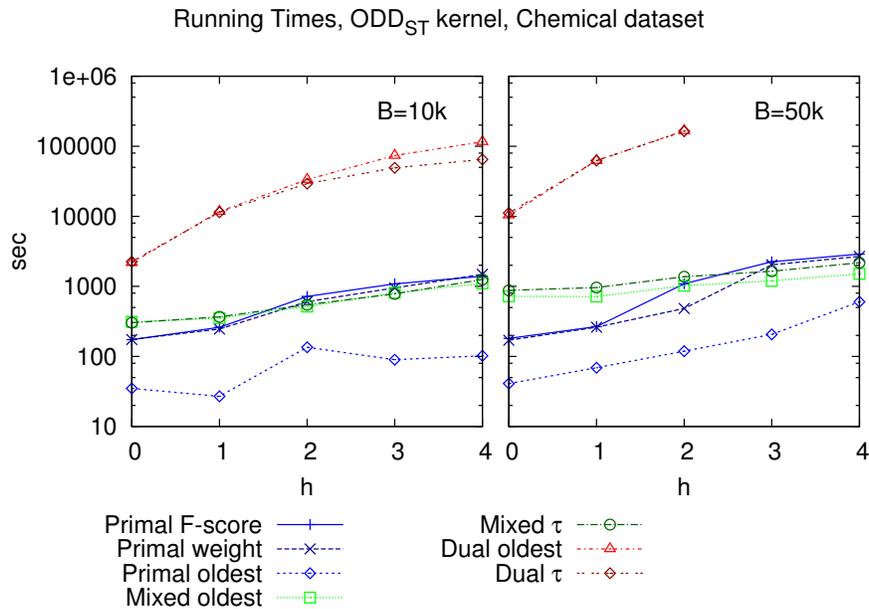


Figure 10: Average computational times of algorithms *Primal*, *Mixed* and *Dual* on the *Chemical* dataset for the ODD_{ST} kernel.

the AUROC measure, while the two on the bottom report running times. One point in a plot represents the AUROC/running time over all *Chemical* dataset for one configuration of the kernel parameters. Note that running times are in logarithmic scale.

Figures 4-5 refer to the FS kernel with *oldest* and *weight* budget management policy, respectively. Note that, by increasing h , the representation in memory of an example does not change for Algorithm 1, whilst it requires more memory for Algorithms 2-3 since the number of features increases. Figures 6-7 refer to the NSPDK kernel (with the same budget values). Each point refers to a combination of the h and d parameters of the kernel (the values are grouped with respect to the h parameter). Figures 8-9 are similar but show the results referring to the ODD_{ST} kernel (values are again grouped with respect to the h parameter). Consider that, on this dataset and with no memory budget constraint on the model, the ODD_{ST} kernel generates a model with a total of 91,467 features with $h = 3$ (the higher the h parameter, the more features are generated). Such number is the size of w ($\|w\|$) and thus the size of the vectorial representation of the model.

Table 1 reports, for each combination of dataset, algorithm, kernel, policy

Table 1: Best *AUROC* value (\pm standard deviation) for each dataset, algorithm, policy, kernel for $10k$ and $50k$ budget values.

kernel	Alg.	Policy	Chemical		Image	
			10k	50k	10k	50k
FS	Primal	weight	.681 \pm .094	.746 \pm .096	.914 \pm .094	.913 \pm .095
		oldest	.626 \pm .092	.659 \pm .093	.917 \pm .092	.918 \pm .090
		F-score	.644 \pm .096	.669 \pm .096	.916 \pm .090	.919 \pm .091
	Mixed	τ	.554 \pm .124	.561 \pm .114	.908 \pm .099	.901 \pm .095
		oldest	.513 \pm .096	.533 \pm .097	.907 \pm .103	.912 \pm .096
	Dual	τ	.547 \pm .127	.582 \pm .115	.907 \pm .093	.906 \pm .094
		oldest	.507 \pm .098	.538 \pm .098	.884 \pm .117	.915 \pm .090
NSPDK	Primal	weight	.707 \pm .091	.762 \pm .092	.907 \pm .095	.907 \pm .095
		oldest	.641 \pm .092	.693 \pm .092	.909 \pm .093	.910 \pm .092
		F-score	.674 \pm .092	.691 \pm .090	.914 \pm .091	.912 \pm .094
	Mixed	τ	.588 \pm .126	.600 \pm .114	.894 \pm .100	.882 \pm .113
		oldest	.519 \pm .101	.532 \pm .102	.899 \pm .106	.907 \pm .091
	Dual	τ	.583 \pm .121	.581 \pm .103	.892 \pm .105	.890 \pm .105
		oldest	.520 \pm .102	.571 \pm .083	.877 \pm .115	.918 \pm .093
ODD _{ST}	Primal	weight	.685 \pm .094	.735 \pm .097	.919 \pm .088	.919 \pm .088
		oldest	.620 \pm .092	.674 \pm .094	.919 \pm .088	.919 \pm .088
		F-score	.661 \pm .098	.693 \pm .097	.919 \pm .088	.919 \pm .088
	Mixed	τ	.572 \pm .125	.574 \pm .125	.909 \pm .093	.905 \pm .107
		oldest	.513 \pm .098	.527 \pm .095	.910 \pm .098	.917 \pm .085
	Dual	τ	.558 \pm .134	.562 \pm .129	.907 \pm .096	.910 \pm .095
		oldest	.504 \pm .097	.518 \pm .097	.883 \pm .120	.907 \pm .098

and budget values $10k$ and $50k$, the best AUROC value among the tested parameters. The table allows to easily compare different policies and different algorithms.

If we consider the *Chemical* dataset, the highest value for *Primal* algorithm is 0.762 (NSPDK, *weight* policy, budget $50k$), while the best AUROC value for Algorithm *Dual* and *Mixed* are 0.583 and 0.600 respectively. Concerning the *F-score* policy of the *Primal* algorithm, since it does not have corresponding policies for *Mixed* and *Dual* algorithms, we decided to omit all *F-score* plots. However, we report the results related to this policy in Table 1. In the *Chemical* dataset, this policy does not improve the predictive performance of the *Primal* algorithm, where the *weight* policy is consistently the

best performing one. Analyzing the plots we can see that the *Primal* algorithm (Algorithm 3) is not only competitive but it always outperforms *Dual* and *Mixed* in both the *weight* and *oldest* policies. Table 1 shows that, practically in all cases, a higher budget increases the classification performance on the *Chemical* dataset, implying that *Dual* and *Mixed* would probably need a significantly higher budget to reach the performances of *Primal* with $B = 10k$.

Unfortunately, setting $B > 50k$ for these algorithms on the *Chemical* dataset is unfeasible because of computational times, as it is possible to see from Figure 10 reports the average time in seconds needed for the three considered algorithms, instantiated with the ODD_{ST} kernel, to process the *Chemical* dataset with $B = 10k$ and $50k$.

The figure shows that there is a clear gap between the computational times of Algorithms *Primal*, *Mixed* and *Dual*. Similar considerations can be drawn for NSPDK and FS kernels. With budget $10k$, the time needed by the *Primal* algorithm to process a single example is on average ($h = \{0 \dots 4\}$) 0.004 seconds, while for the *Dual* algorithm the required time is 0.2 seconds. The gap grows when setting the budget to $50k$. In this case the *Primal* algorithm needs on average 0.006 seconds, while for the *Dual* algorithm already with $h = 0$ the required time per example is 0.05 seconds (almost ten times slower than *Primal*), with $h=1$ it is 0.39 seconds. With $h = 3$ and 4 the experiments did not complete in 48 hours, meaning that the processing of each example required more than 1 second on average. The *Mixed* algorithm has computational times similar to the *Primal* ones, but with considerably worse predictive performance.

To summarize the results, Figure 11 shows, for each algorithm, the classification performance in relation to the running time, for budget $10k$ and $50k$. The plots report one point for each algorithm, kernel and parameters combination. We can see that the *Primal* algorithm has many points in the upper/left part of the plot, meaning that it is able to achieve high predictive performances in a relatively small amount of computational time. *Mixed* and *Dual* algorithms are all over the lower part of the plot, meaning that they have worse predictive performances and higher running times than *Primal*.

4.3.2. Experiments on the Image Datasets

The same experimental setting described for the *Chemical* dataset is replicated here for the *Image* dataset. Figures 12-17 show, for each set of corresponding management policies, the performance of the kernels with respect

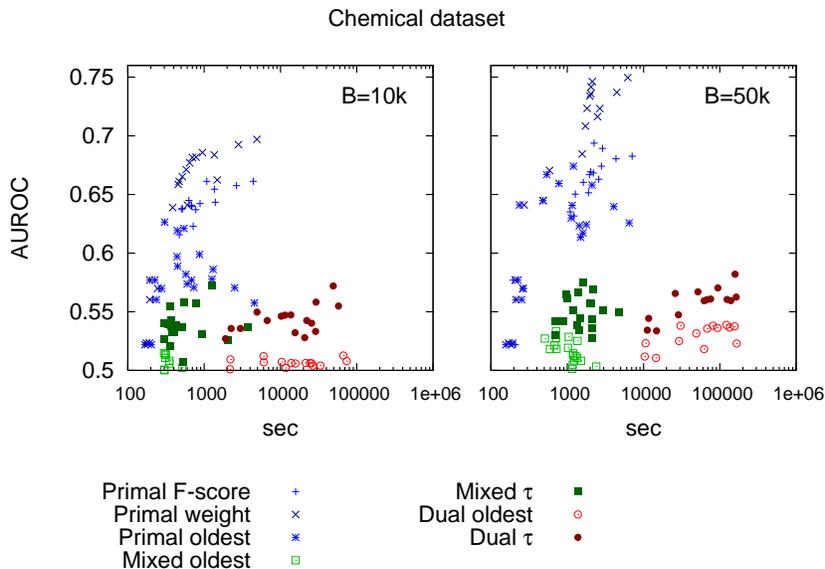


Figure 11: Comparison among computational times and AUROC of algorithms *Primal*, *Mixed* and *Dual* on the Image dataset with budget $10k$ and $50k$ for all the considered policies and kernels.

to their parameters. We tested different values for the budget size, ranging from $1k$ to $100k$. In Figure 12 we can see that, for small budget values, the *Primal* algorithm is the best performing one with the *oldest* budget management policy. When the budget grows (i.e. for $B = 100k$) *Mixed* and *Dual* perform slightly better than *Primal*. Figure 13, referring to the same kernel with *weight* policy, depicts a similar scenario. In this case, *Primal* performs slightly better than *Dual* and *Mixed* in all the considered budget sizes.

In Figures 14 and 15 we started from a budget value of $B = 2.5k$, since the NSPDK generates more features than FS (as detailed in Section 2.2). When considering the *oldest* policy, *Primal* performs best for budget values up to $10k$. In the case of *weight* policy, *Primal* always performs better than *Dual* and *Mixed*. More in general, it is possible to see that the performance of *Dual* and *Mixed* increase proportionally to the budget, while *Primal* performs best with budget $10k$, thus its performance do not improve if more budget is available (note nonetheless that the performance do not decrease significantly). Apparently, in the case of FS and NSPDK kernels, the classification performances of the different algorithms depend critically on the budget size. Figure 16 analyzes the situation with ODD_{ST} kernel and *oldest*

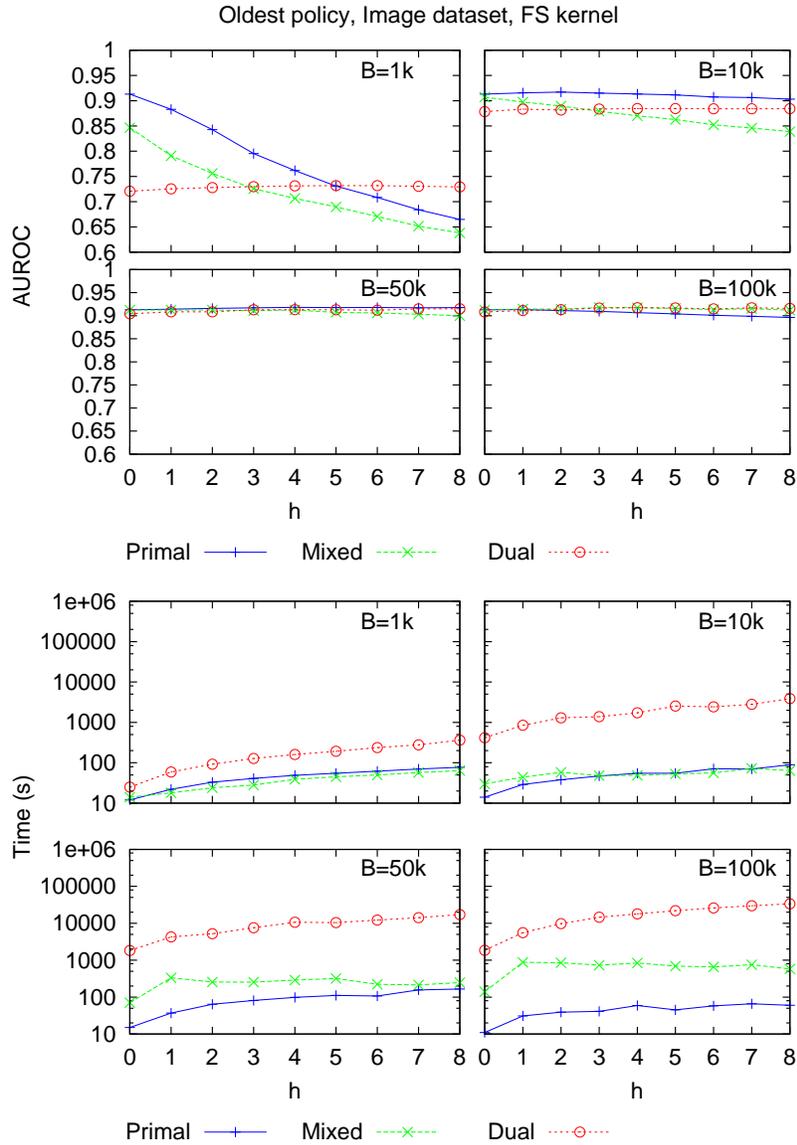


Figure 12: Average AUROC value computed over all stream instances for memory budgets $B = 1k$ (top left), $B = 10k$ (top right), $B = 50k$ (bottom left) and $B = 100k$ (bottom right) for algorithms *Primal*, *Mixed* and *Dual* with respect to the values of the *FS* kernel parameter. Below the first set of plots there is a second one with the corresponding running times. Plots refer to the *Image* dataset and the *oldest* budget maintenance policy.

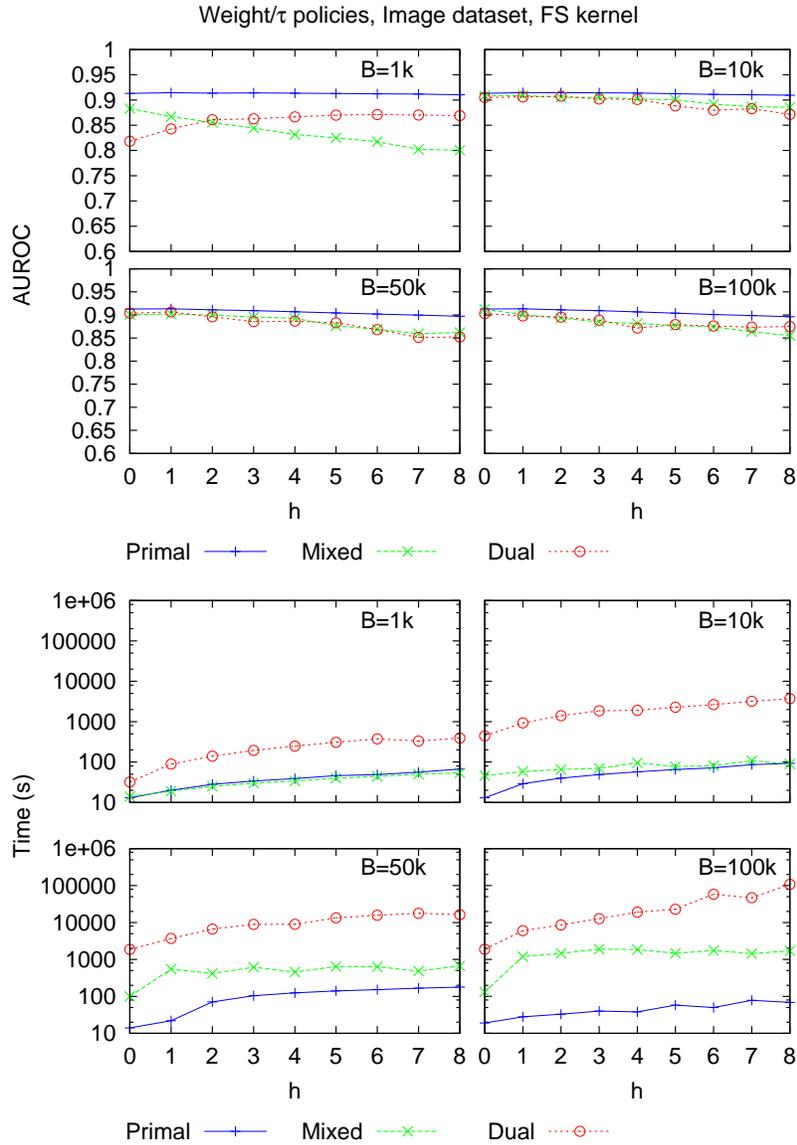


Figure 13: Average AUROC value computed over all stream instances for memory budgets $B = 1k$ (top left), $B = 10k$ (top right), $B = 50k$ (bottom left) and $B = 100k$ (bottom right) for algorithms *Primal*, *Mixed* and *Dual* with respect to the values of the *FS* kernel parameter. Below the first set of plots there is a second one with the corresponding running times. The plots refer to the *Image* dataset and the *weight*/ τ budget maintenance policies.

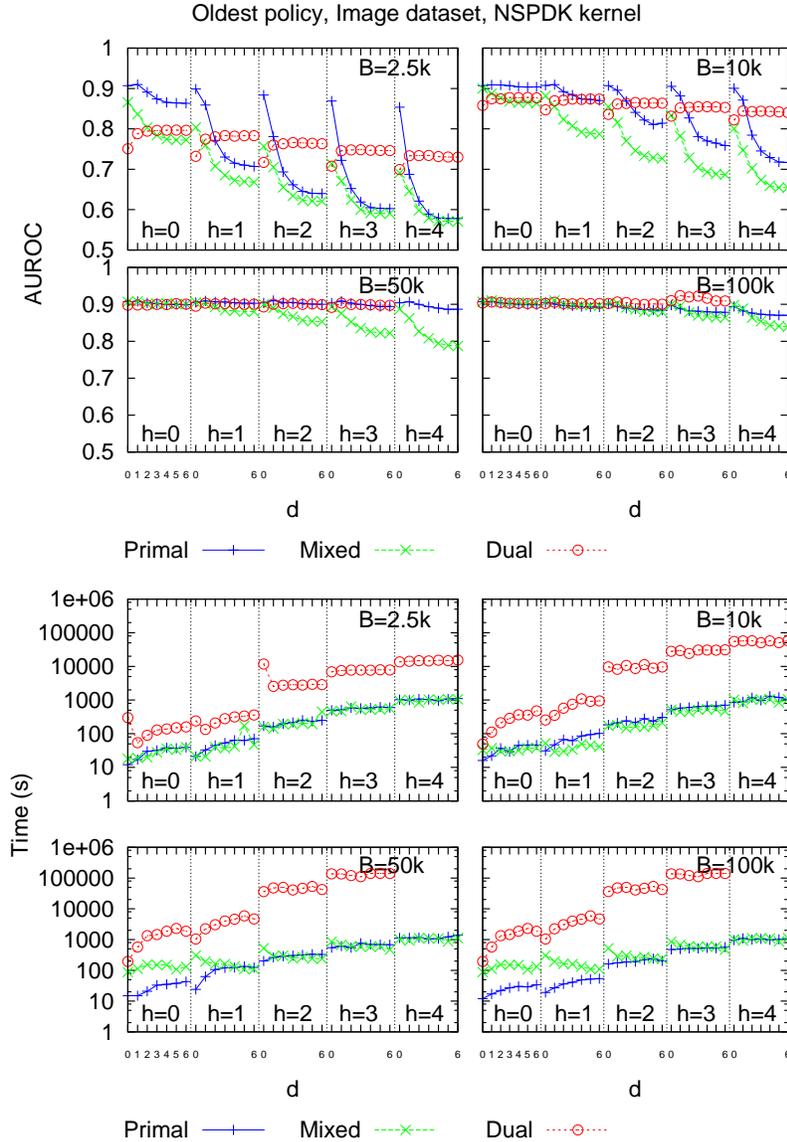


Figure 14: Average AUROC value computed over all stream instances for memory budgets $B = 2.5k$ (top left), $B = 10k$ (top right), $B = 50k$ (bottom left) and $B = 100k$ (bottom right) for algorithms *Primal*, *Mixed* and *Dual* with respect to the values of the *NSPDK* kernel parameters. Below the first set of plots there is a second one with the corresponding running times. The plots refer to the *Image* dataset and the *oldest* budget maintenance policy. Missing values indicate that the corresponding execution has not terminated in 48 hours.

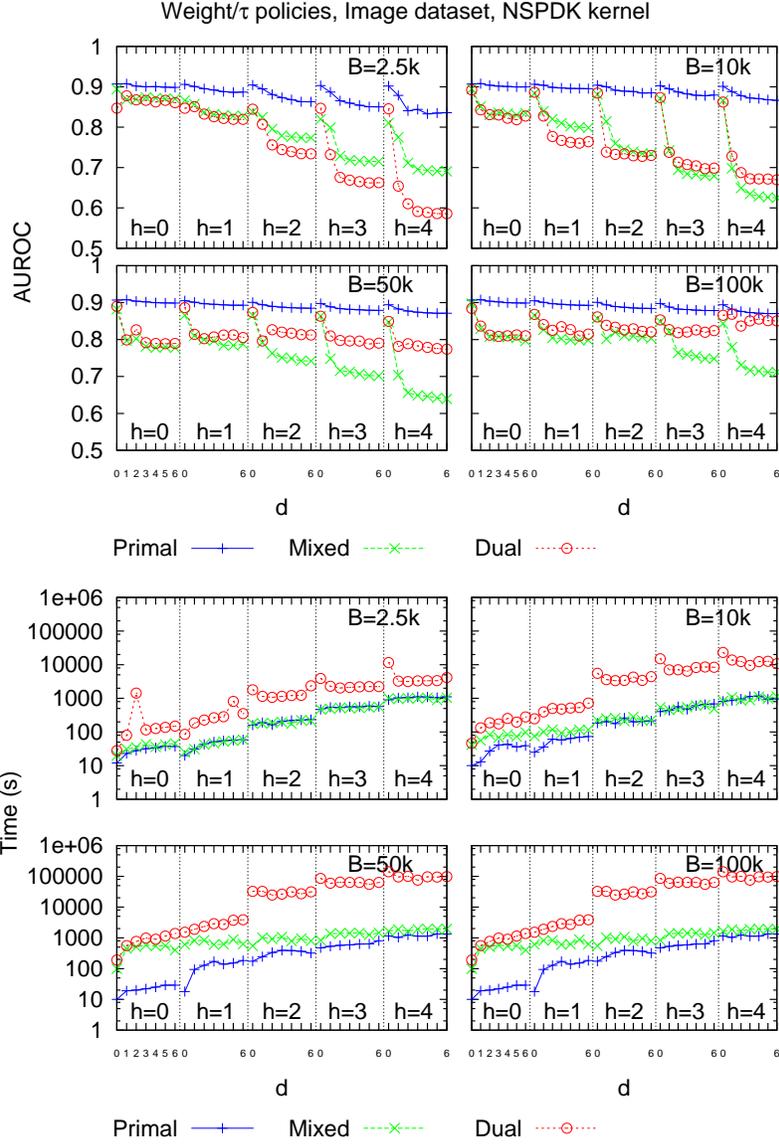


Figure 15: Average AUROC value computed over all stream instances for memory budgets $B = 2.5k$ (top left), $B = 10k$ (top right), $B = 50k$ (bottom left) and $B = 100k$ (bottom right) for algorithms *Primal*, *Mixed* and *Dual* with respect to the values of the *NSPDK* kernel parameters. Below the first set of plots there is a second one with the corresponding running times. The plots refer to the *Image* dataset and the *weight*/ τ budget maintenance policies.

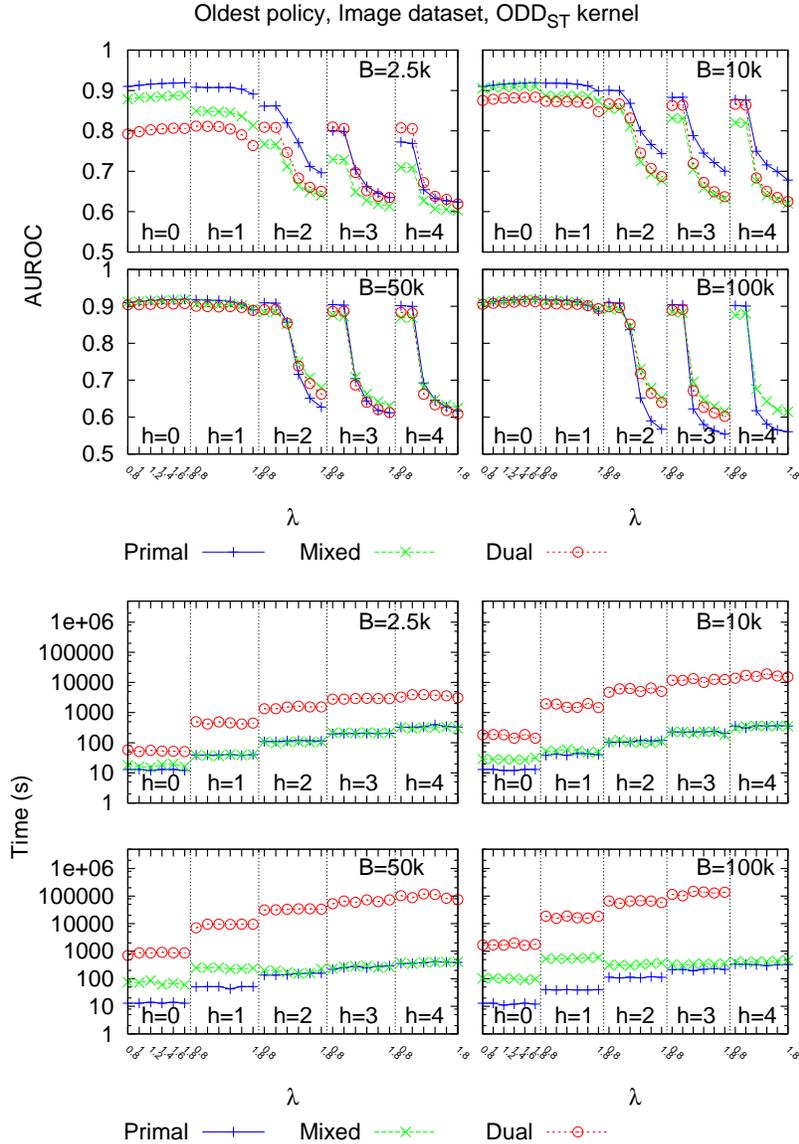


Figure 16: Average AUROC value computed over all stream instances for memory budgets $B = 2.5k$ (top left), $B = 10k$ (top right), $B = 50k$ (bottom left) and $B = 100k$ (bottom right) for algorithms *Primal*, *Mixed* and *Dual* with respect to the values of the the ODD_{ST} kernel parameters. Below the first set of plots there is a second one with the corresponding running times. The plots refer to the *Image* dataset and the *oldest* budget maintenance policy. Missing values indicate that the corresponding execution has not terminated in 48 hours.

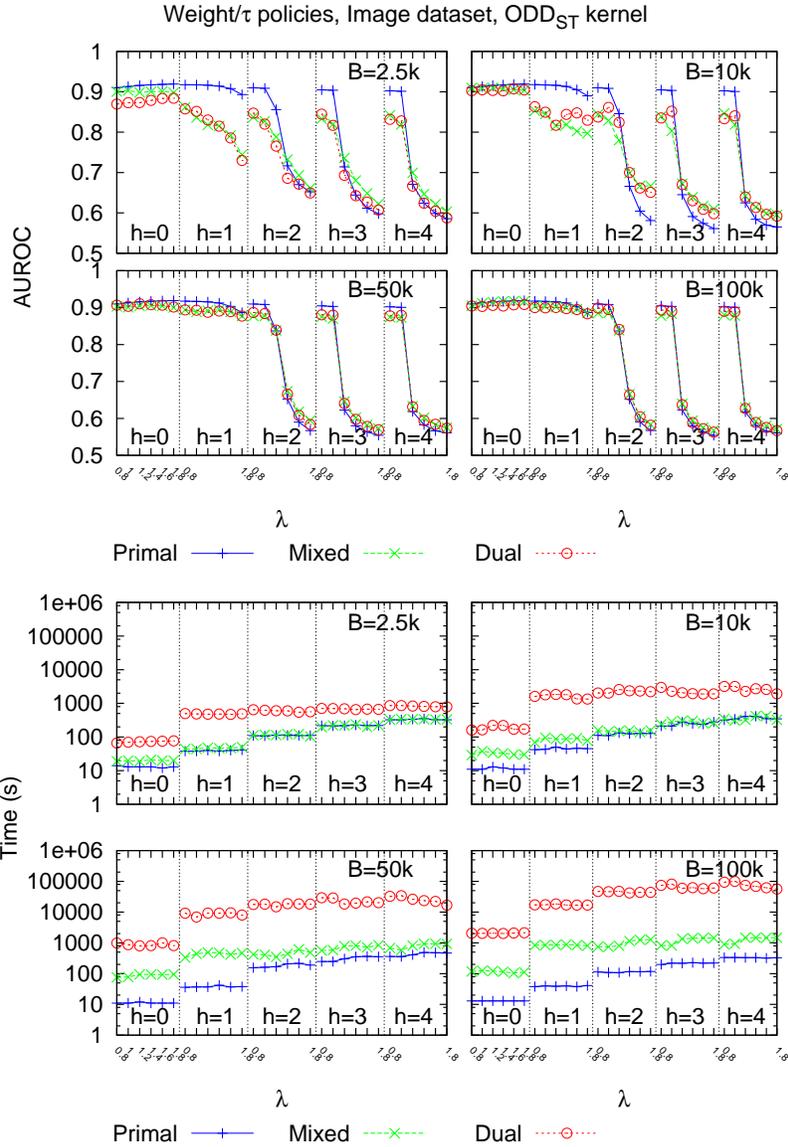


Figure 17: Average AUROC value computed over all stream instances for memory budgets $B = 2.5k$ (top left), $B = 10k$ (top right), $B = 50k$ (bottom left) and $B = 100k$ (bottom right) for algorithms *Primal*, *Mixed* and *Dual* with respect to the values of the the ODD_{ST} kernel parameters. Below the first set of plots there is a second one with the corresponding running times. The plots refers to the *Image* dataset and the *weight*/ τ budget maintenance policies.

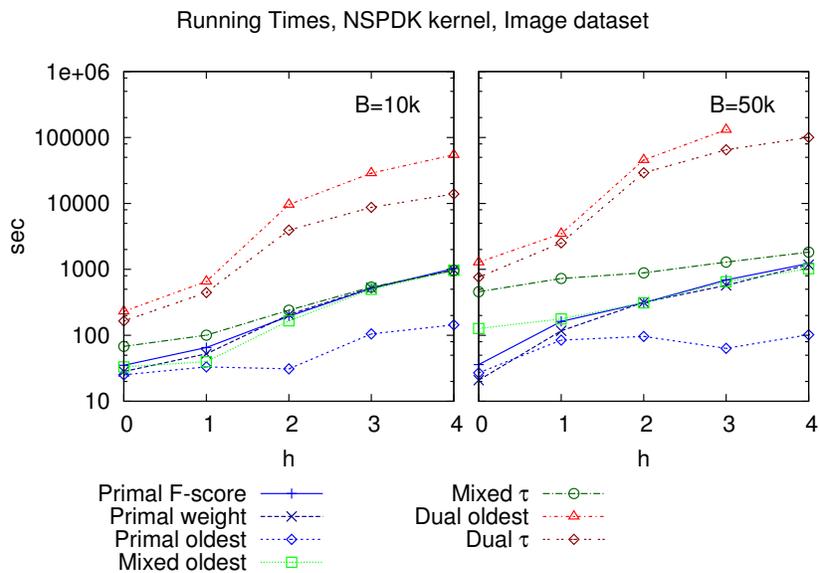


Figure 18: Average computational times of algorithms *Primal*, *Mixed* and *Dual* on the *Image* dataset for the NSPDK kernel.

policy. Also in this case, *Primal* algorithm is the better performing one with every budget value. However, with higher budgets, the other algorithms show comparable performances. Also in this case, the higher the budget the better the predictive performances of *Mixed* and *Dual*. The scenario is similar when considering the *weight*/ τ policies in Figure 17.

The running times of the different kernels on the *Image* dataset are in general lower with respect to the *Chemical* one. Figure 18 reports the running time required by the FS kernel with budget 10,000. As for the *Chemical* dataset the *Primal* and *Mixed* algorithms are way faster than the *Dual* algorithm.

Figure 19 shows the predictive performance in relation to the computational time required from the different algorithms in the *Image* dataset. The *Primal* algorithm is the fastest, with some points at the leftmost margin of the plots. Also from a predictive performance point of view, we see that the algorithm with the highest AUROC is *Primal* for both budget values. With $B = 50k$ the *Mixed* and *Dual* algorithms achieve similar performances, although with a higher runtime.

To summarize, given a budget management policy, under a certain budget size *Primal* algorithm is the best performing one, and over that size *Dual* and

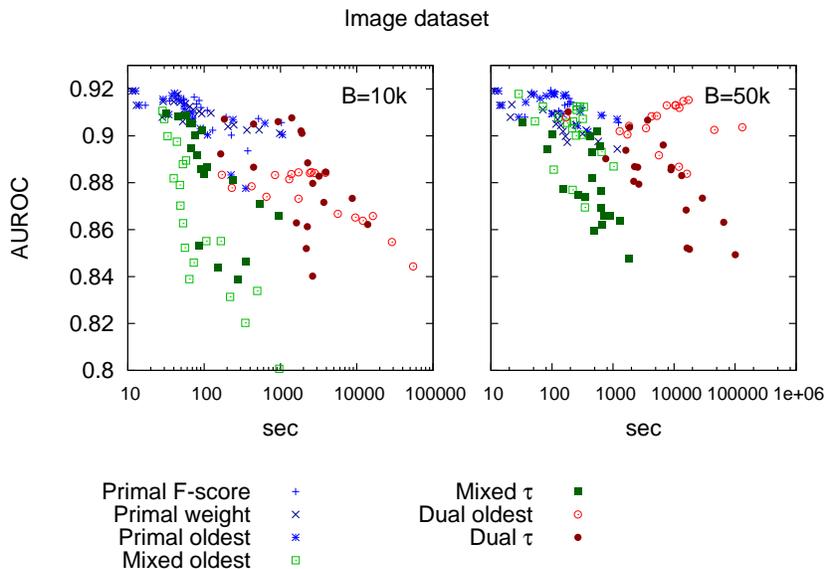


Figure 19: Comparison among computational times and AUROC of algorithms *Primal*, *Mixed* and *Dual* on the Image dataset with budget $10k$ and $50k$ for all the considered policies and kernels.

Primal (and in some cases *Mixed*) show very similar performances. However, there is a significant difference in the computational times required by the different algorithms, with *Primal* and *Mixed* being considerably faster than *Dual*.

4.3.3. Discussion

We can draw some final remarks concluding our experimental analysis. First it is worth to point out that our analysis refers only to those kernels which allow for an explicit feature space representation. Such kernels are only a subset of the existing graph kernels. However, they are the ones currently having state-of-the-art predictive performances. While the *Dual* algorithm can represent more compactly the model than the *Primal* approach when the feature space associated to the kernel is very large, this implies a loss in efficiency when computing the score for a new graph: the kernel value between the input graph and all the graphs in the model have to be computed from scratch. As the values of Figures 10 and 18 indicate, that makes the application of the *Dual* algorithm to graph streams practically infeasible, especially when strict time constraints have to be satisfied. The *Mixed*

algorithm is able to significantly speed up the score computation by storing the explicit feature space representation of each graph in the model. As a consequence, the size of the model may increase significantly, thus reducing the total number of graphs that can be kept in it: *Dual* algorithm is able to store in memory approximately 250 graphs of the chemical datasets with budget 10,000, while *Mixed* algorithm only 100 graphs. On the contrary, *Primal* algorithm keeps in the model only the most informative features, and thus it is able to retain information of all graphs inserted in the model while preserving a very good efficiency. According to our experiments, there is a budget value which determines whether the *Primal* or the other approaches are preferable. While such threshold value can be observed in our experiments for the *Image* dataset, due to the inefficiency of *Dual* and *Mixed*, we were not able to identify it for the *Chemical* dataset (where *Primal* always outperforms the other approaches).

5. Conclusions and Future Work

In this work we analyzed the trade-off between efficiency and efficacy of various versions of online margin kernel perceptron algorithms when dealing with graph streams and under the assumption of fixed memory budgets. One of them efficiently exploits the explicit representation of the feature space (via hash tables) of different state-of-the-art graph kernels recently defined in literature.

Experimental results on real-world datasets show that, under a threshold budget size, working in feature space is preferable both in terms of classification performance and running times. In a future work we will investigate the dependency between such budget value and the size of the feature space associated to the kernel, the policy for pruning the model and the nature of the dataset.

6. Acknowledgments

This work was supported by the University of Padova under the strategic project *BIOINFOGEN*.

References

- [1] R. Klinkenberg, Learning drifting concepts: Example selection vs. example weighting, *Intell. Data Analy.* 8 (2004) 281–300.

- [2] M. Eskandari, S. Hashemi, A graph mining approach for detecting unknown malwares, *J. Vis. Lang. Comput.* 23 (3) (2012) 154–162.
- [3] C. Alippi, S. Ntalampiras, M. Roveri, A Cognitive Fault Diagnosis System for Distributed Sensor Networks, *IEEE Trans. Neural Netw. Learning Syst.* 24 (8) (2013) 1213–1226.
- [4] C. C. Aggarwal, On Classification of Graph Streams, in: *SDM*, 652–663, 2011.
- [5] J. Gibert, E. Valveny, H. Bunke, Embedding of graphs with discrete attributes via label frequencies, *International Journal of Pattern Recognition and Artificial Intelligence* 27 (03).
- [6] A. M. Bianucci, A. Micheli, A. Sperduti, A. Starita, Application of Cascade Correlation Networks for Structures to Chemistry, *Applied Intelligence* 12 (1-2) (2000) 117–146.
- [7] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 1 edn., 2000.
- [8] T. Hofmann, B. Schölkopf, A. J. Smola, Kernel methods in machine learning, *The Annals of Statistics* 36 (3) (2008) 1171–1220.
- [9] G. Da San Martino, A. Sperduti, Mining Structured Data, *IEEE Computational Intelligence Magazine* 5 (1) (2010) 42–49.
- [10] G. Da San Martino, N. Navarin, A. Sperduti, A Tree-Based Kernel for Graphs, in: *Proceedings of the Twelfth SIAM International Conference on Data Mining*, 975–986, 2012.
- [11] F. Aioli, G. Da San Martino, A. Sperduti, A. Moschitti, Fast On-line Kernel Learning for Trees, in: *Proceedings of the 2006 IEEE Conference on Data Mining*, IEEE Computer Society, Los Alamitos, CA, USA, 787–791, 2006.
- [12] J. Kivinen, A. J. Smola, R. C. Williamson, Online learning with kernels, *IEEE Transactions on Signal Processing* 52 (8) (2004) 2165–2176.

- [13] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, Y. Singer, Online Passive-Aggressive Algorithms, *Journal of Machine Learning Research* 7 (2006) 551–585.
- [14] F. Orabona, J. Keshet, B. Caputo, Bounded Kernel-Based Online Learning, *Journal of Machine Learning Research* 10 (2009) 2643–2666.
- [15] T. Gärtner, P. Flach, S. Wrobel, On graph kernels: Hardness results and efficient alternatives, *Lecture notes in computer science* (2003) 129–143.
- [16] K. M. Borgwardt, H.-P. Kriegel, Shortest-Path Kernels on Graphs, in: *Proceedings of the Fifth IEEE International Conference on Data Mining*, IEEE Computer Society, 74–81, 2005.
- [17] P. Mahé, J. Vert, Graph kernels based on tree patterns for molecules, *Machine Learning* 75 (1) (2009) 3–35.
- [18] N. Shervashidze, K. Borgwardt, Fast subtree kernels on graphs, in: Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22*, Curran Associates, Inc., 1660–1668, 2009.
- [19] F. Costa, K. D. Grave, Fast Neighborhood Subgraph Pairwise Distance Kernel, in: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, June 21-24, 2010, Haifa, Israel, 255–262, 2010.
- [20] H. Kashima, K. Tsuda, A. Inokuchi, Marginalized Kernels Between Labeled Graphs., in: T. Fawcett, N. Mishra (Eds.), *ICML*, AAAI Press, 321–328, 2003.
- [21] K. M. Borgwardt, N. N. Schraudolph, S. Vishwanathan, Fast Computation of Graph Kernels, in: B. Schölkopf, J. Platt, T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19*, MIT Press, 1449–1456, 2007.
- [22] M. Heinonen, N. Välimäki, V. Mäkinen, J. Rousu, Efficient Path Kernels for Reaction Function Prediction, in: *BIOINFORMATICS 2012 - Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*, Vilamoura, Algarve, Portugal, 1 - 4 February, 2012., 202–207, 2012.

- [23] N. Shervashidze, S. V. N. Vishwanathan, T. H. Petri, K. Mehlhorn, K. M. Borgwardt, Efficient Graphlet Kernels for Large Graph Comparison, in: D. van Dyk, M. Welling (Eds.), Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 5 of JMLR: Workshop and Conference Proceedings, CSAIL, Clearwater Beach, Florida, USA, 488–495, 2009.
- [24] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, *Psychological review* 65 (6) (1958) 386–408.
- [25] G. Cavallanti, N. Cesa-Bianchi, C. Gentile, Tracking the best hyperplane with a simple budget Perceptron, *Mach. Learn.* 69 (2-3) (2007) 143–167.
- [26] O. Dekel, S. S. Shwartz, Y. Singer, The Forgetron: A kernel-based perceptron on a fixed budget, in: *Advances in Neural Information Processing Systems 18*, MIT Press, Cambridge, MA, 1342–1372, 2006.
- [27] K. Crammer, J. Kandola, Y. Singer, Online Classification on a Budget, in: S. Thrun, L. Saul, B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA, 225–232, 2003.
- [28] Z. Wang, S. Vucetic, Online Passive-Aggressive Algorithms on a Budget, *Journal of Machine Learning Research - Proceedings Track 9* (2010) 908–915.
- [29] J. Langford, L. Li, T. Zhang, Sparse Online Learning via Truncated Gradient, *Journal of Machine Learning Research* 10 (2009) 777–801.
- [30] J. C. Duchi, Y. Singer, Efficient Online and Batch Learning Using Forward Backward Splitting, *Journal of Machine Learning Research* 10 (2009) 2899–2934.
- [31] Y.-W. Chen, C.-J. Lin, Combining SVMs with Various Feature Selection Strategies, in: I. Guyon, M. Nikravesh, S. Gunn, L. Zadeh (Eds.), *Feature Extraction*, vol. 207 of Studies in Fuzziness and Soft Computing, Springer Berlin Heidelberg, 315–324, 2006.

- [32] B. Russell, A. Torralba, LabelMe: a database and web-based tool for image annotation, *International journal of Computer Vision* 77 (1-3) (2008) 157–173.
- [33] P. Su, R. L. S. Drysdale, A comparison of sequential Delaunay triangulation algorithms, in: *Proceedings of the eleventh annual symposium on Computational geometry - SCG*, ACM Press, New York, New York, USA, 61–70, 1995.

An Empirical Study on Budget-Aware Online Kernel Algorithms for Streams of Graphs: Supplementary material.

The Balanced Accuracy (BAC) that is defined as the arithmetic mean of sensitivity and specificity, or the average accuracy obtained on either class:

$$BAC = \frac{1}{2} \left(\frac{tn}{tn + fp} + \frac{tp}{fn + tp} \right),$$

where tp , tn , fp and fn are, respectively, true positive, true negative, false positive and false negative predictions. The results reported here adopt this measure.

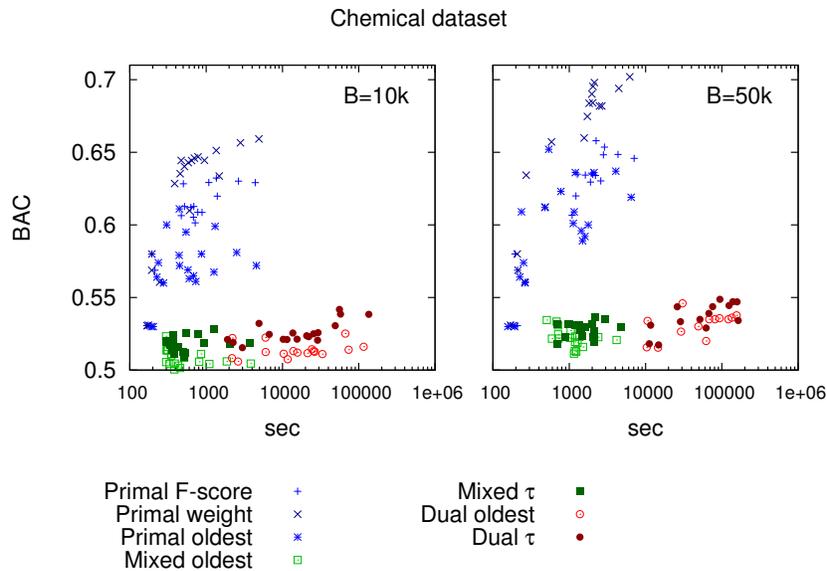


Figure 1: Comparison among computational times and BER of algorithms *Primal*, *Mixed* and *Dual* on the Image dataset with budget $10k$ and $50k$ for all the considered policies and kernels.

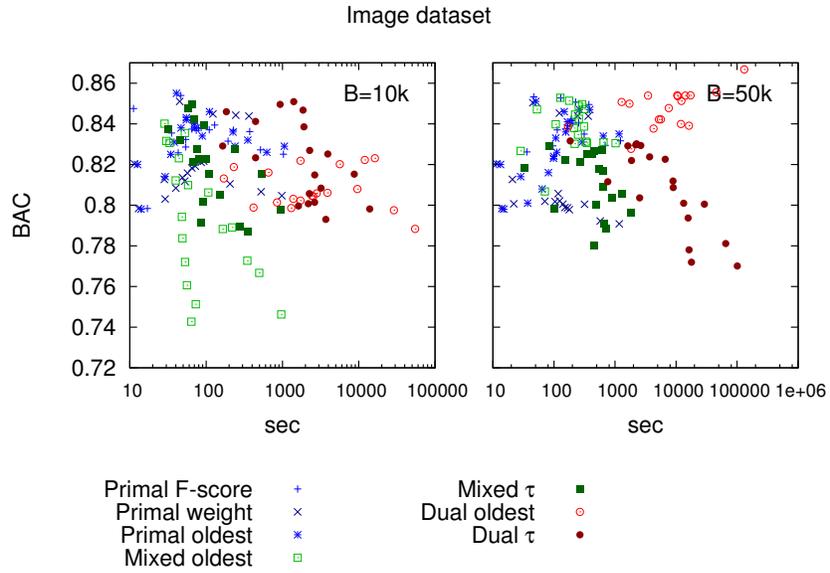


Figure 2: Comparison among computational times and BER of algorithms *Primal*, *Mixed* and *Dual* on the Image dataset with budget $10k$ and $50k$ for all the considered policies and kernels.

Table 1: Best BAC value for each dataset, algorithm, policy, kernel and budget.

kernel	Algorithm	Policy	Chemical		Image				
			10k	50k	1k	2.5k	10k	50k	100k
FS	Primal	weight	0.646	0.698	0.823	0.827	0.821	0.802	0.801
		oldest	0.600	0.623	0.798	0.825	0.843	0.841	0.801
		F-score	0.612	0.635	0.798	0.819	0.839	0.842	0.838
	Mixed	τ	0.519	0.531	0.808	0.839	0.849	0.827	0.811
		oldest	0.513	0.530	0.738	0.794	0.831	0.849	0.857
	Dual	τ	0.525	0.548	0.782	0.827	0.850	0.823	0.812
		oldest	0.514	0.537	0.604	0.701	0.806	0.854	0.854
	NSPDK	Primal	weight	0.663	0.704	0.807	0.814	0.812	0.812
oldest			0.611	0.652	0.805	0.831	0.842	0.835	0.814
F-score			0.637	0.652	0.791	0.828	0.837	0.837	0.835
Mixed		τ	0.524	0.535	0.796	0.825	0.827	0.805	0.785
		oldest	0.525	0.537	0.721	0.780	0.830	0.852	0.856
Dual		τ	0.538	0.544	0.800	0.819	0.829	0.811	0.792
		oldest	0.529	0.548	0.586	0.696	0.818	0.866	0.849
ODD _{ST}		Primal	weight	0.644	0.684	0.850	0.850	0.851	0.850
	oldest		0.595	0.636	0.823	0.850	0.855	0.851	0.851
	F-score		0.629	0.657	0.826	0.843	0.853	0.853	0.853
	Mixed	τ	0.527	0.536	0.823	0.843	0.837	0.829	0.840
		oldest	0.516	0.524	0.761	0.812	0.840	0.847	0.851
	Dual	τ	0.541	0.534	0.797	0.820	0.845	0.831	0.838
		oldest	0.516	0.520	0.638	0.748	0.823	0.849	0.856