Sequence Classification in the Jensen-Shannon Embedding

Jérôme Callut JEROME.CALLUT@UCLOUVAIN.BE
Pierre Dupont PIERRE.DUPONT@UCLOUVAIN.BE
Department of Computing Science and Engineering (INGI),
Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium
UCL Machine Learning Group (MLG)

Marco Saerens SAERENS@ISYS.UCL.AC.BE
Faculty of Economical, Social and Political Sciences (ESPO), School of Management (IAG),
Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium
UCL Machine Learning Group (MLG)

Abstract

This paper presents a novel approach to the supervised classification of structured objects such as sequences, trees and graphs, when the input instances are characterized by probability distributions. Distances between distributions are computed via the Jensen-Shannon (JS) divergence, which offers several advantages over the L_2 distance or the Kullback-Leibler divergence. The JS divergence induces an embedding of the distributions into a real Hilbert space. A general approach is proposed here to derive a positive definite kernel from any conditionally negative definite (CND) distance, the JS divergence being a particular case of interest. We show how to compute the dot product in the embedding induced by a CND distance, based solely on the distance matrix between training points, and we detail how new points can be added to this embedding. The JS kernel is applied to sequence classification problems. Two kinds of empirical distributions are considered: (i) the N-gram distributions and (ii) the distributions of the First Passage Times (FPT) between occurrences of substrings. Experimental results on DNA splicing junction detection and protein function prediction illustrate that ...

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

This paper is concerned with a supervised classification problem in which the instances are sequences defined over a discrete alphabet. Practical applications of this task range from the recognition of boundaries between introns and exons in DNA sequences to musical pieces classification. In this context, we propose to chararcterize the sequences by empirical probability distributions: the N-gram (contiguous subsequences of length N) distributions and the First Passage Times (FPT) distributions. Sequences defined by these features are classified using a new kernel based on the Jensen-Shannon divergence.

A naive approach to compare two discrete distributions is to represent them as vectors and to apply the classical L_2 distance. However, the L_2 distance is based on differences between probabilities while it is often more relevant to use probability (log-)ratios. In this context, measures based on the Shannon entropy such as the Kullback-Leibler divergence (Lin, 1991) and the Jensen-Shannon divergence (Lin, 1991) are commonly used. The Jensen-Shannon (JS) divergence has several advantages over the Kullback-Leibler divergence. In particular, in this paper we exploit the fact that the JS divergence is the square of a true metric distance and that it is a conditionally negative definite (CND) function. According to the Schoenberg theorem (Schoenberg, 1938), the JS divergence induces an embedding of the distributions in a real Hilbert space. The Jensen-Shannon kernel is defined as the dot product in this embedding. Schölkopf (Schölkopf, 2000; Schölkopf & Smola, 2002) has shown that conditionally positive definite (CPD) kernels (for instance obtained by taking the

opposite of CND distances) can directly be used in translation invariant learning algorithms. The present paper describes a more general technique based on the multidimensional scaling theory (MDS) (Cox & Cox, 2001; Mardia et al., 1979) to compute a true positive definite (PD) kernel from any CND distance. In contrast to CPD kernels, the resulting kernel can be used in any kernel-based algorithm, no matter whether the learning method is affected by translations. The kernel corresponds to a centered dot product in the space induced by the distance. The origin of the Hilbert space is located at the centroid of the configuration of points. This centering is relevant for the JS divergence as this function is relative to the mean of the probability distributions. We also propose a matrix-based technique to classify new points in the induced JS embedding. One of the main advantages of the JS kernel is that it is parameter-free, avoiding the need to tune extra hyperparameters. A previous work (Chan et al., 2004) has already proposed a kernel, called here the modified Gaussian kernel, based on the Jensen-Shannon divergence. It is obtained by replacing the squared Euclidean distance by the JS divergence in the Gaussian kernel. This kernel, however, does not correspond to the dot product in the JS embedding and, as the Gaussian kernel, it requires to tune a hyperparameter. An alternative approach (Jebara & Kondor, 2003) relies on building a generative model for each instance. The kernel value for two instances is obtained by integrating the product of the two corresponding generative models. This technique is more complex than the JS kernel since it requires to select an appropriate kind of generative models and a practical way to integrate the product of these densities.

The proposed kernel is applied to solve discrete sequence classification problems. We consider two kinds of features, i.e. of empirical distributions, extracted from the sequences. The first are the joint distributions of N-grams. These features have been successfully applied to real-world applications (Shawe-Taylor & Cristianini, 2004), however, they can fail to model long-range dependencies or complex time-dependent dynamics. Rather than considering local features like N-grams, an alternative approach relies on dynamical features in the sequences. In this perspective, we focus on the time distribution between occurrences of substrings in the sequences. More precisely, given a pair of substrings (v, w), we are looking at the number of steps taken to observe the next occurrence of w after having observed v. The distribution of these measures forms the First Passage Time (FPT) dynamics of a sequential process with respect to the feature (v, w). A previous work (Callut & Dupont, 2006) proposed to model the FPT with phase-type distributions. Two classifiers based on phase-type distributions were presented: (i) a maximum *a posteriori* classifier and (ii) a SVM with a marginalization kernel. These techniques are time consuming since the number of substring pairs can be large and for each pair a phase-type distribution has to be estimated. In this paper, rather than building a probabilistic model of the FPT, we propose a fully discriminative technique by applying the JS kernel on the empirical FPT extracted from the sequences.

The rest of this paper is organized as follows. Section 2 reviews the Jensen-Shannon divergence. Section 3 presents the Schoenberg theorem and shows how to compute the dot product (i.e. the kernel) in the embedding induced by a CND distance when only distances between the training points are known. Section 4 focus on sequence classification problems using the JS kernel with N-gram and FPT features. Finally, section 5 shows experimental results obtained on the two following tasks: (i) DNA splicing junction detection and (ii) protein function prediction.

2. The Jensen-Shannon Divergence

The Jensen-Shannon divergence is a function which measures the distance between two distributions (Lin, 1991). While comparing samples defined by their empirical distribution, a larger JS divergence indicates that they are more likely to have been drawn from different source distributions (i.e. from different classes). Let $\mathcal{P}$ denote the space of all probability distributions defined over a discrete set of events[1] Ω. The JS divergence is a function $\mathcal{P} \times \mathcal{P} \to \mathbb{R}$ defined by:

$$D_{JS}(P_1, P_2) = H(M) - \frac{1}{2}H(P_1) - \frac{1}{2}H(P_2)$$

where $P_1, P_2 \in \mathcal{P}$ are two distributions, $M = \frac{1}{2}(P_1 + P_2)$ and $H(P) = -\sum_{e\in\Omega} P[e] \log P[e]$ is the Shannon entropy. Since the Shannon entropy $H(.)$ is a concave function, the Jensen inequality implies that the JS divergence is non-negative, furthermore it is bounded by 1 (Lin, 1991). The JS divergence is closely related to the Kullback-Leibler (KL) divergence, a classical measure to compare two distributions. Indeed,

$$D_{JS}(P_1, P_2) = \frac{1}{2}D_{KL}(P_1, M) + \frac{1}{2}D_{KL}(P_2, M)$$

where $D_{KL}(P_1, P_2) = \sum_{e\in\Omega} P_1[e] \log \frac{P_1[e]}{P_2[e]}$. The JS divergence can be thought of as a symmetrized and

[1] More rigourously, the distributions are defined on a measurable space $(\Omega, \mathcal{B})$ where $\mathcal{B}$ is a σ-algebra defined on Ω.

smoothed variant of the KL divergence as it is relative to the mean of the distributions. Furthermore, a natural generalization of the JS divergence (i.e. the generalized JS divergence (Lin, 1991)) allows one to compare more than two distributions and to parametrize the relative importance of distributions with weights. In addition, the square root of the JS divergence, denoted here by $\sqrt{D_{JS}(.,.)}$, enjoys the properties of a true distance metric.

In section 3, we show that $\sqrt{D_{JS}(.,.)}$ induces an embedding of the distributions in a real Hilbert space $\mathcal{F}$. That is, $\sqrt{D_{JS}(P_1,P_2)} = \|\phi(P_1) - \phi(P_2)\|$ where $\phi : \mathcal{P} \to \mathcal{F}$ is a function that maps a probability distribution in a Hilbert space $\mathcal{F}$. The Jensen-Shannon kernel is defined as the dot product $\langle \phi(.), \phi(.) \rangle$ in $\mathcal{F}$. A fundamental property used to prove the existence of the embedding (see section 3) is the notion of *conditionally negative definite* function. In the sequel, $\mathcal{X}$ will denote a general input space (for the JS divergence $\mathcal{X} = \mathcal{P}$). For the sake of simplicity, a positive/negative definite function actually denotes a semi-positive/semi-negative definite function in this paper.

Definition 1 *A symmetric function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that, for all $n \in \mathbb{N}$, $c_i \in \mathbb{R}$ and for all $x_i \in \mathcal{X}$:*

$$\sum_{i=1}^{n} c_i = 0 \Rightarrow \sum_{i,j=1}^{n} c_i c_j f(x_i, x_j) \leq 0$$

is called a conditionally negative definite (CND) function.

The (usual) definition of a negative definite function is obtained by removing the condition $\sum_{i=1}^{m} c_i = 0$. Therefore, any negative definite function is also a CND function. It has recently been shown that the JS divergence is CND (Topøe, 2003).

3. Embedding of a CND distance into a Hilbert space

In section 2, it was shown that the square of the JS divergence is a CND distance metric. Schölkopf (Schölkopf, 2000; Schölkopf & Smola, 2002) has mentioned that conditionally positive definite (CPD) kernels[2] are "as good" as positive definite (PD) kernels when used in translation invariant algorithms such as SVM and kernel PCA. We propose here a general technique based on the multidimensional scaling theory (MDS) to compute a PD kernel from a CND squared distance. The resulting kernel is also valid for kernel-based algorithms affected by translations (e.g. the cosine classifier). This section first reviews the Schoenberg theorem (Schoenberg, 1938) which states that a distance $d(.,.)$ defines an embedding of the points into a Hilbert space $\mathcal{F}$ *if and only if* $d(.,.)^2$ is CND. The distance between any pair of points x_i, x_j in this embedding is exactly the distance $d(x_i, x_j)$.

Theorem 1 (Schoenberg (Schoenberg, 1938)) *Let $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a distance function between two elements of a set $\mathcal{X}$. There is a Hilbert space embedding of $(\mathcal{X}, d(.,.))$* if and only if *$d(.,.)^2$ is a conditionally negative definite function.*

In the rest of this section, we show how to compute a PD dot product matrix relative to a finite set of points when only their CND squared distances are known. This result, stated in theorems 2 and 3, constitutes a constructive proof for the "if" part of the Schoenberg theorem when a finite set of points is considered. Since our technique is completely matrix-based (i.e. no analytic expression of the kernel is provided, nor needed), we also show how to embed new points in $\mathcal{F}$. These derivations can be made for any distance function such that its square is CND and in particular for $\sqrt{D_{JS}(.,.)}$.

For the sake of completeness, let us show that, if there exists a Hilbert space embedding, the square of the distance is CND (i.e. the "only if" part). Let $\mathcal{F}$ denotes a Hilbert space and $\phi(x_1), \ldots, \phi(x_n)$ a set of $n \in \mathbb{N}$ points mapped into this space. The squared distance between two points $\phi(x_i), \phi(x_j)$ is defined as $\|\phi(x_i) - \phi(x_j)\|^2 = \|\phi(x_i)\|^2 + \|\phi(x_j)\|^2 - 2\langle \phi(x_i), \phi(x_j) \rangle$. We shall show that it is CND: $\sum_{i=1}^{n} c_i = 0 \Rightarrow \sum_{i,j=1}^{n} c_i c_j \|\phi(x_i) - \phi(x_j)\|^2 = -2 \sum_{i,j=1}^{n} c_i c_j \langle \phi(x_i), \phi(x_j) \rangle = -2 \| \sum_{i=1}^{n} c_i \phi(x_i) \|^2 \leq 0$.

3.1. Derivation of a kernel matrix from a CND distance

In this section, we construct a PD kernel matrix K for a finite set of $n \in \mathbb{N}$ points given their CND distance matrix. An interpretation of this kernel in the embedding is given in section 3.2. Let $\{x_1, \ldots, x_n\}$ be a set of $n \in \mathbb{N}$ points lying in the input space $\mathcal{X}$ and let A denote an $n \times n$ matrix with $a_{ij} = -\frac{1}{2} d(x_i, x_j)^2$. The kernel matrix K is defined as

$$K = HAH \quad \text{with } H = I - \frac{E}{n} \tag{1}$$

where I is a $n \times n$ identity matrix and E is an $n \times n$ matrix with each element being equal to 1. H is a

[2] A CPD kernel $k(.,.)$ can be obtained from a CND distance $d(.,.)$ by defining $k(x_i, x_j) = -\frac{1}{2} d(x_i, x_j)$ for all $x_i, x_j \in \mathcal{X}$.

projection matrix and is called the centering matrix (Mardia et al., 1979). It projects the vectors into a $n-1$ dimensional space orthogonal to the column vector 1, which means that $(Hz)^T 1 = 0$ for all $z \in \mathbb{R}^n$. Let us show that the matrix K is PD.

Theorem 2 *Let $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a distance function between two elements of a set $\mathcal{X}$ and $x_1, \ldots, x_n \in \mathcal{X}$. Let A denote a $n \times n$ matrix with $a_{ij} = -\frac{1}{2} d(x_i, x_j)^2$ and let H denote a $n \times n$ matrix defined as $H = I - \frac{E}{n}$. Then, the matrix $K = HAH$ is PD if $d(.,.)^2$ is CND.*

Proof.

If $d(.,.)$ is CND, the following inequalites holds

$$\begin{aligned} -2\,(Hz)^T A(Hz) &\leq 0 \quad \forall \boldsymbol{z} \in \mathbb{R}^n \\ \iff z^T(HAH)z &\geq 0 \quad \forall \boldsymbol{z} \in \mathbb{R}^n \end{aligned}$$

which explicitely defines $K = HAH$ as a PD matrix.

■

According to the Mercer theorem (Mercer, 1909), K is a dot product matrix for a set of n points embedded into a Hilbert space: $k_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$ for all $1 \leq i, j \leq n$. Let us show that the distance between two points x_i, x_j in this embedding is $d(x_i, x_j)$.

Theorem 3 *Let K be a PD kernel matrix defined as in equation 1. Then, the distance between two points x_i, x_j in the embedding induced by K is $d(x_i, x_j)$.*

Proof.

The squared distance between two points x_i, x_j with $1 \leq i, j \leq n$ in the embedding relative to K is

$$\begin{aligned} \|\phi(x_i) - \phi(x_j)\|^2 &= k_{ii} + k_{jj} - 2k_{ij} \\ &= (e_i - e_j)^T HAH(e_i - e_j) \\ &= (e_i - e_j)^T (I - \frac{E}{n}) A (I - \frac{E}{n})(e_i - e_j) \\ &= (e_i - e_j)^T A (e_i - e_j) = -2a_{ij} \end{aligned}$$

where e_i is an $n \times 1$ vector with all elements being equal to 0 except for the i-th being equal to 1. ■

To sum up, the Schoenberg theorem states that any distance $d(.,.)$, in particular the JS divergence, having a CND square induces an embedding in a Hilbert space. This result is proved here constructively for a finite set of points by deriving a positive definite kernel matrix K and by showing that the distance in the embedding induced by K is $d(.,.)$.

3.2. Interpretation of the kernel in the embedding

In this section, an interpretation of the kernel defined in equation 1 is given. Starting from a CND distance matrix, we show how to compute the relative dot product matrix which turns out to be K. This derivation makes use of a technique used in MDS (Cox & Cox, 2001; Mardia et al., 1979) for computing a dot product matrix from a distance matrix, adapted here to work in the feature space. Since a CND distance $d(.,.)$ induces a Hilbert space embedding, one can write $d(x_i, x_j)^2 = \|\phi(x_i) - \phi(x_j)\|^2 = \|\phi(x_i)\|^2 + \|\phi(x_j)\|^2 - 2\langle \phi(x_i), \phi(x_j) \rangle$ (2)

Let us observe that while the distance is translation invariant, this is not the case for the dot product. To overcome this indetermination, the origin of $\mathcal{F}$ is placed at the centroid of the n points. This amounts to consider the translated mapping $\varphi : \mathcal{X} \to \mathcal{F}$ such that $\varphi(x) = \phi(x) - \frac{1}{n} \sum_{k=1}^n \phi(x_k)$. It should be noticed that this centering is very natural for the JS divergence as this function is relative to the mean of the probability distributions. Hence, the kernel operates the following computation in the original embedding:

$$\begin{aligned} \langle \varphi(x_i), \varphi(x_j) \rangle &= \langle \phi(x_i), \phi(x_j) \rangle - \frac{1}{n} \langle \phi(x_i), \sum_{k=1}^n \phi(x_k) \rangle \\ &- \frac{1}{n} \langle \phi(x_j), \sum_{k=1}^n \phi(x_k) \rangle + \frac{1}{n^2} \langle \sum_{k=1}^n \phi(x_k), \sum_{k=1}^n \phi(x_k) \rangle \end{aligned}$$

By using the fact that $\sum_{k=1}^n \varphi(x_k) = 0$ and by summing on the index i and on the indexies i, j in the formula (2) (which is also valid for $\varphi(.)$) one respectively obtains the identities (3) and (4):

$$\begin{aligned} \|\varphi(x_i)\|^2 &= \frac{1}{n} \sum_{k=1}^n \langle \varphi(x_k), \varphi(x_k) \rangle \\ &- \frac{1}{n} \sum_{k=1}^n \|\varphi(x_i) - \varphi(x_k)\|^2 \end{aligned} \quad (3)$$

$$\frac{2}{n} \sum_{k=1}^n \langle \varphi(x_k), \varphi(x_k) \rangle = \frac{1}{n^2} \sum_{k,l=1}^n \|\varphi(x_k) - \varphi(x_l)\|^2 \quad (4)$$

Substituting these identities in the formula (2), the dot product $\langle \varphi(x_i), \varphi(x_j) \rangle$ becomes :

$$\begin{aligned} &\langle \varphi(x_i), \varphi(x_j) \rangle = -\frac{1}{2} (\|\varphi(x_i) - \varphi(x_j)\|^2 \\ &= \frac{1}{n} \sum_{k=1}^n \|\varphi(x_i) - \varphi(x_k)\|^2 + \frac{1}{n} \sum_{k=1}^n \|\varphi(x_j) - \varphi(x_k)\|^2 \\ &- \frac{2}{n^2} \sum_{k,l=1}^n \|\varphi(x_k) - \varphi(x_l)\|^2) \end{aligned} \quad (5)$$

Interestingly, the dot product in the Hilbert space is now completely computable using only distances[3] be-

[3] Since the distance is translation invariant, $\|\varphi(x_i) - \varphi(x_j)\|^2 = \|\phi(x_i) - \phi(x_j)\|^2$.

tween the points. Equation (5) can be rewritten in matrix form as $K = HAH$ which is exactly the kernel matrix defined in section 3.1.

3.3. Adding new points in the embedding

Let us see now how to add a new point, that is, to compute the dot product between the existing points of the training set and a new point to be classified. We assume that we have computed the distances between the new point and all the training set samples or the support vectors for a SVM classifier, according to our CND distance; the vector containing the distances between the n training set samples and the new $(n+1)$th point will be denoted by δ. Thus δ_i, the ith element of δ, is the square distance between x_i and the new point x_{n+1} in the input space.

We are seeking for a new embedded point $\varphi(x_{n+1})$ at the respective squared distance $d(x_i, x_{n+1})^2 = \delta_i$ from each training samples $\varphi(x_i)$. A similar problem occurs in the theory of MDS and has been studied by Gower in (Gower, 1968); this section is largely inspired by this work. This new point needs not be computed explicitly in the embedding but the dot product between the new point and the training set samples is computed from the distances vector δ and the kernel matrix K. The distance between the new point and the training samples must be preserved in the embedding space:

$$\begin{aligned} \delta_i &= \|\varphi(x_i) - \varphi(x_{n+1})\|^2 = \|\varphi(x_i)\|^2 + \|\varphi(x_{n+1})\|^2 \\ &- 2\langle\varphi(x_i), \varphi(x_{n+1})\rangle \\ &= k_{ii} + \|\varphi(x_{n+1})\|^2 - 2\langle\varphi(x_i), \varphi(x_{n+1})\rangle \end{aligned} \tag{6}$$

where k_{ij} denotes the element (i, j) of the kernel matrix K, computed on the training set. Let us first compute $\|\varphi(x_{n+1})\|^2$ by summing the previous equation over i:

$$\sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} k_{ii} + n\|\varphi(x_{n+1})\|^2 \tag{7}$$

since the $\varphi(x_i)$ are centered. From this last equation, we can easily deduce $\|\varphi(x_{n+1})\|^2$

$$\|\varphi(x_{n+1})\|^2 = \frac{1}{n}[\sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} k_{ii}] \tag{8}$$

By replacing $\|\varphi(x_{n+1})\|^2$ in equation (6) and extracting $\langle\varphi(x_i), \varphi(x_{n+1})\rangle$, we finally obtain

$$\langle\varphi(x_i), \varphi(x_{n+1})\rangle = \frac{1}{2}[(k_{ii} - \frac{1}{n}\sum_{i=1}^{n} k_{ii}) - (\delta_i - \frac{1}{n}\sum_{i=1}^{n} \delta_i)] \tag{9}$$

which corresponds to the dot product between the new point and point i from the training set in the embedding space. If we denote by k^{new} the column vector containing the $\langle\varphi(x_i), \varphi(x_{n+1})\rangle$ as elements, we obtain in matrix form

$$k^{new} = \frac{H}{2}[diag(K) - \delta] \tag{10}$$

where $diag(K)$ is a column vector made of the diagonal of K. This result can easily be generalized to a set of new points. If Δ is a matrix containing the distances between the training samples (as rows) and the new points (as columns), and K^{new} contains the dot products between the training samples (as rows) and the new points (as columns),

$$K^{new} = \frac{H}{2}[diag(K)1^T - \Delta] \tag{11}$$

4. Application to sequence classification

In the present section we describe how the JS kernel can be used to solve a supervised classification problem in which the instances are sequences defined over a discrete alphabet Σ and the labels belong to a set $\mathcal{Y}$. Given a set of n examples $\{(s_1, y_1), \ldots, (s_n, y_n)\}$ where $s_i \in \Sigma^*$ is a sequence and $y_i \in \mathcal{Y}$ its label (or its class), one wants to estimate a function $f : \Sigma^* \to \mathcal{Y}$ for predicting the label of new sequences. Two kinds of empirical distributions are considered for computing the JS kernel: (i) the N-gram distributions and (ii) the distributions of the First Passage Times (FPT) between occurrence of substrings. The sequence classification problem is solved using support vector machines (SVM) with the JS kernel. SVM is a binary classifier computing the maximal margin hyperplane in the space induced by a kernel. Once the SVM has been trained, new sequences are classified by mapping them in the JS Hilbert space (see section 3.3) and by looking at which side of the hyperplane they lie.

The first considered features extracted are the empirical probability distributions of N-grams (i.e. contiguous subsequences of length N) for a fixed N. Given a sequence s, the (unconditional) probability of an N-gram w is defined as $\hat{P}(w) = \frac{C(w,s)}{|s|-N+1}$ where $|s|$ is the length of the sequence s and $C(w, s)$ is the number of times the N-gram w occurs in s. It should be pointed out that these distributions are defined over the set of events $\Omega = \Sigma^N$ and that the unseen N-grams have a null probability (no smoothing is applied). The JS distance is computed for each sequence pair (s_i, s_j) and the kernel matrix is obtained as shown in section 3.1. These features are closely related to the features of the

p-spectrum kernel (Shawe-Taylor & Cristianini, 2004) (p. 347). This kernel directly computes the dot product between N-grams counts extracted from sequences while the JS kernel operates on relative N-gram frequencies.

The second kind of features are the First Passage Times (FPT) between occurrence of substrings.

Definition 2 *Given a sequence s defined on an alphabet Σ and two substrings $v, w \in \Sigma^+$. For each occurrence of v in s, the first passage time to w is defined as the finite number of steps taken before observing the next occurrence of w. The* first passage times *from v to w in s is a multiset defined as the first passage times to w for all occurrences of v in s.*

For instance, let us consider the alphabet $\Sigma = \{a, b\}$, the sequence $s = aababba$ and the events $v = ab$ and $w = ba$. The value of the pair (ab, ba) is $\phi_{(ab,ba)}(s) = \{3, 1\}$. Let us note that the step count starts after the last character of v and it does not take the length of w into account. If one consider substrings up to length N, the potential number of features is bounded by $(\sum_{i=1}^{N} |\Sigma|^i)^2 \in \mathcal{O}(|\Sigma|^{2N})$, where $|\Sigma|$ denotes the alphabet size. However the number of features observed in a pair of training sequences is often by far below this bound. It can be shown that the number of features observed in a pair of sequences is upper bounded by $L^2.N^2$, where L is the length of the longest sequence. For a given substring pair (v, w), the empirical FPT distribution in a sequence s is obtained by computing the relative frequency of each time in $\phi_{(v,w)}(s)$. Such a distribution is defined over $\Omega = \mathbb{N}$ and unobserved FPT have a null probability. The JS kernel is then applied to these distributions and summed up over all considered pairs. One should notice that when computing $k(s, s')$, if a pair does not appear in s and s', the JS divergence equals 0 since by convention $0 \log 0 = 0$. For the same reason, if a pair only appears in one of the two sequences, the JS divergence is equal to 0.5.

5. Experiments

This section is intended to investigate the three following points: (i) what are the benefits of the JS kernel with respect to standard sequence kernels such as the p-spectrum kernel? (ii) what is the improvement of FPT features over N-gram features? and (iii) what is the advantage of the JS kernel with respect to the modified Gaussian kernel applied on the same distributions? To answer these questions, we consider two sequence classification tasks: (i) DNA splicing site detection (`Splice` dataset) and (ii) protein function prediction (`Kinase` database). The `Splice` dataset[4] is made of windows of 60 symbols from DNA sequences containing intro-exon (IE) or exon-intron (EI) boundaries or neither of them. We restrict here our attention to binary classification by considering sequences labeled either EI or IE. The class priors are equal and the data set contains 1481 sequences. The `Kinase` dataset is based on the families of human kinases and contains 290 protein sequences belonging to four functional classes. The four classes contain respectively 210, 69, 10 and 1 sequence(s). Given a kinase sequence, the objective is to predict one of the four kinase subfamilies. All experiments were carried out using a 10-fold stratified[5] cross-validation (CV). Results are averaged over the 10 folds and a 95% confidence interval is provided. For the SVM-based method, one fold has been held out as a validation set to tune the regularization constant C and the α parameter in the case of the modified Gaussian kernel.

TO BE COMPLETED

References

Andorf, C., Silvescu, A., D.Dobbs, & Honavar, V. (2004). Learning classifiers for assigning protein sequences to gene ontology (go) functional families. *Proceedings of the Fifth International Conference On Knowledge Based Computer Systems (KBCS).*

Bengio, Y., & Frasconi, P. (1995). Diffusion of context and credit information in markovian models. *Journal of Artificial Intelligence Research, 3*, 223–244.

Callut, J., & Dupont, P. (2005). Inducing hidden markov models to model long-term dependencies. *16th European Conference on Machine Learning (ECML)* (pp. 513–521). Porto, Portugal: Springer Verlag.

Callut, J., & Dupont, P. (2006). Sequence discrimination using phase-type distributions. *17th European Conference on Machine Learning (ECML)* (pp. 78–89). Berlin, Germany: Springer Verlag.

Chan, A. B., Vasconcelos, N., & Moreno, P. J. (2004). *A family of probabilistic kernels based on information divergence* (Technical Report). Statistical Visual Computing Lab, Department of Electrical and Computer Engineering, University of California, San Diego. SVCL-TR 2004/01.

Cox, T., & Cox, M. (2001). *Multidimensional scaling, 2nd ed.* Chapman and Hall.

[4] `Splice` is available from the UCI repository.

[5] Stratified CV tends to reproduce the same class priors in the folds as in the complete dataset.

Gower, J. (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika*, *55*, 582–585.

Jebara, T., & Kondor, R. I. (2003). Bhattacharyya expected likelihood kernels. *COLT* (pp. 57–71).

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. *International Conference on Acoustic, Speech and Signal Processing* (pp. 181–184). Detroit, Michigan.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Trans. Information Theory*, *37*, 145–151.

Mardia, K., Kent, J., & Bibby, J. (1979). *Multivariate analysis*. Academic Press.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc. London*, *A 209*, 415–446.

Ralaivola, L., Swamidass, S. J., Saigo, H., & Baldi, P. (2005). Graph kernels for chemical informatics. *Neural Netw.*, *18*, 1093–1110.

Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.*, *44*, 522–536.

Schölkopf, B. (2000). The kernel trick for distances. *NIPS* (pp. 301–307).

Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. Cambridge: MIT Press.

Schreiber, F., & Schwöbbermeyer, H. (2005). Frequency concepts and pattern detection for the analysis of motifs in networks. *Transactions on Computational Systems Biology*, *3 (LNBI 3737)*, 89–104.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.

Topøe, F. (2003). *Jensen-shannon divergence and norm-based measures of discrimination and variation* (Technical Report). Department of Mathematics, University of Copenhagen.

Yakhnenko, O., Silvescu, A., & Honavar, V. (2005). Discriminatively trained markov model for sequence classification. *ICDM* (pp. 498–505).