kernels assume a Euclidean distance to compare feature vectors. Symbolic kernels, in contrast, have been less commonly used in the published literature. We use the following distance-based kernel functions in our analysis.

1. **Linear (Euclidean) Kernel Function for Real-valued Features:** This is the most commonly used kernel function for linear SVMs and other kernel-based algorithms. Let *x* and *z* be two feature vectors. Then the function

$$\mathrm{K}_l(x,z)=x^T z \tag{1}$$

defines a Euclidean distance-based kernel function. KPCA with linear kernel function reduces to the standard PCA. The linear kernel trivially follows Mercer's condition for kernel validity, that is, the matrix comprising pairwise kernel function values for any finite subset of feature vectors selected from the feature space is guaranteed to be positive semidefinite.

2. **Hamming Kernel Function for Nominal Features:** Let the number of features be *N*. Let *x* and *z* be two feature vectors, that is, *N*-dimensional symbolic vectors from a finite, symbolic feature space X_N. Then the function

$$\mathrm{K}_h(x,z)=N-\sum_{i=1}^{N}\delta(x_i,z_i) \tag{2}$$

where $\delta(x_i,z_i)=0$ if $x_i=z_i$ and 1 otherwise, defines a Hamming distance–based kernel function and follows the Mercer's condition for kernel validity (Aradhye & Dorai, 2002). The equality, $x_i=z_i$, refers to symbol/label match.

3. **Cityblock Kernel Function for Discrete Features:** Using the preceding notation, the function

$$\mathrm{K}_c(x,z)=\sum_{i=1}^{N}M_i-|x_i-z_i| \tag{3}$$

where the *i*th feature is Mi-ary, defines a Cityblock distance–based kernel function and follows the Mercer's condition for kernel validity (Aradhye & Dorai, 2002).

4. **Edit Kernel Function for Stringlike Features:** We define an edit kernel between two strings as

$$\mathrm{K}_e(x,z)=\max(len(x),len(z))-E(x,z) \tag{4}$$

where $E(x,z)$ is the edit distance between the two strings, defined conventionally as the minimum number of change, delete, and insert operations required to convert one string into another, and $len(x)$ is the length of string *x*. In theory, Edit distance does *not* obey Mercer validity, as has been recently proved by Cortes and coworkers (Cortes, Haffner, & Mohri, 2002, 2003). However, empirically, the Kernel matrices generated by the edit kernel are often positive definite, justifying the practical use of Edit distance–based kernels.

Hybrid Multimodal Kernel

Having defined these four basic kernel functions for different modalities of features, we are now in a position to define a multimodal kernel function that encompasses all types of common multimedia features. Let any given feature vector *x* be comprised of a real-valued feature set x_r, a nominal feature set x_h, a discrete-valued feature set x_c, and a string-style feature x_e, such that $x=[x_r \quad x_h \quad x_c \quad x_e]$. Then, because a linear combination of valid kernel functions is a valid kernel function, we define

$$\mathrm{K}_m(x,z)=\alpha\mathrm{K}_l(x_r,z_r)+\beta\mathrm{K}_h(x_h,z_h)+\gamma\mathrm{K}_c(x_c,z_c)+\delta\mathrm{K}_l(x_e,z_e) \tag{5}$$

where $\mathrm{K}_m(x,z)$ is our multimodal kernel and α, β, γ, and δ are constants. Such a hybrid kernel can now be seamlessly used to analyze multimodal feature vectors that have real and symbolic values, without imposing any artificial integer mapping of symbolic labels and further obtaining the benefits of analyzing disparate data together as one. The constants α, β, γ, and δ can be determined in practice either by a knowledge-based analysis of the relative importance of the different types of features or by empirical optimization.

Example Multimedia Application: Videotext Postprocessing

Video sequences contain a rich combination of images, sound, motion, and text. Videotext, which refers to superimposed text on images and video frames, serves as an important source of semantic information in video streams, besides speech, close caption, and visual content in video. Recognizing text superimposed on video frames yields important information such as the identity of the speaker, his/her location, topic under discussion, sports scores, product names, associated shopping data, and so forth, allowing for automated content description,