Kernel Machine Classification Using Universal Embeddings

Petros T. Boufounos and Hassan Mansour
Mitsubishi Electric Research Laboratories,
Cambridge, MA 02139, USA, {*petrosb,mansour*}@merl.com

Abstract: Visual inference over a transmission channel is becoming an important problem in a variety of applications where, low latency and bit-rate consumption are often critical performance metrics, making data compression necessary. In this paper, we examine feature compression for support vector machine (SVM)-based inference using quantized randomized embeddings.

Universal Embeddings for Kernel Machines

We consider universal embeddings [1], namely transformations of the form $\phi(\mathbf{x}) = Q(\mathbf{A}\mathbf{x} + \mathbf{e})$, where $\mathbf{A} \in \mathbb{R}^{M \times N}$ is a randomly generated matrix with i.i.d. standard normal entries, $\mathbf{e} \in \mathbb{R}^M$ is a random dither with elements drawn from an i.i.d. distribution uniform in $[0, \Delta]$, $Q(y)$ is a non-monotonic scalar quantizer applied element-wise to its vector input, mapping y to 1 if $y \in [2k, 2k+1)$ and to -1 otherwise, Δ is a scaling parameter, and $\mathbf{x} \in \mathbb{R}^N$ is the vector being embedded—typically a feature vector or a signal to be classified. Universal embeddings have been shown to satisfy

$$g\left(\left\|\mathbf{x} - \mathbf{x}'\right\|_2\right) - \tau \leq d_H\left(\phi(\mathbf{x}), \phi(\mathbf{x}')\right) \leq g\left(\|\mathbf{x} - \mathbf{x}\|_2\right) + \tau, \tag{1}$$

with overwhelming probability, where τ decreases as $1/\sqrt{M}$, $d_H(\cdot,\cdot)$ is the Hamming distance of the embedded signals and $g(d)$ is the map

$$g(d) = \frac{1}{2} - \sum_{i=0}^{+\infty} \left(\pi(i+1/2)\right)^{-2} e^{-\left(\frac{\pi(2i+1)d}{\sqrt{2}\Delta}\right)^2} \approx \frac{d}{\Delta}\sqrt{\frac{2}{\pi}}, \text{if } d \leq \frac{\Delta}{2}\sqrt{\frac{\pi}{2}}, \text{ or 0.5 otherwise.} \tag{2}$$

We demonstrate that SVM kernels based on universal embeddings are very good approximations of radial basis function (RBF) kernels commonly used in classification. Thus, embedding features to a lower dimensional space is equivalent to using the SVM kernel trick with a kernel that approximates an RBF kernel.

Proposition. *Let $\phi(\mathbf{x}) : \mathbb{R}^N \to \{-1, 1\}^M$ be a mapping function defined as above, with $\mathbf{q} = \phi(\mathbf{x})$. The kernel function $K(\mathbf{x}, \mathbf{x}')$ given by $K(\mathbf{x}, \mathbf{x}') = \frac{1}{2M}\mathbf{q}^T\mathbf{q}'$ is shift invariant and approximates the radial basis function $K(\mathbf{x}, \mathbf{x}') \approx \frac{1}{2} - g\left(\|\mathbf{x} - \mathbf{x}'\|_2\right)$, with $g(d)$, as defined in (2). Furthermore, this RBF approximates the Gaussian RBF.*

Our experimental results on an 8-class image database using histogram-of-gradients (HOG) features demonstrate that universal embeddings achieve 50% rate reduction over scalar quantization of the feature vectors, while maintaining the same inference performance.

References

[1] P. T. Boufounos and S. Rane, "Efficient coding of signal distances using universal quantized embeddings," in *Proc. Data Compression Conference (DCC)*, Snowbird, UT, March 20-22 2013.