Ranked Leakage Sensitive Hashing

Rohith Prakash

Consider n streams of time series describing the utilization traces of n resources, each at some (fixed) granularity. Denote the space of a time series as S.

Define a function d:

$$\begin{aligned} d_{a,b} &: S^2 \to [0,1] \\ (x[a:b], y[a:b]) &\mapsto [0,1] \end{aligned} \tag{1}$$

where d maps a portion of two time series to a real number in $[0,1]$ that represents the distance between the partial time series.

Additionally, we require d to be a metric (and satisfy the triangle inequality):

$$\begin{aligned} d(x,y) = 0 &\Leftrightarrow x = y \\ d(x,y) &= d(y,x) \geq 0 \\ d(x,y) &\leq d(x,z) + d(z,y) \end{aligned} \tag{2}$$

Out of the box Dynamic Time Warping (DTW) does not satisfy the triangle inequality, so it cannot be used. We may attempt to leverage an approximate lower-bound DTW [1] as well as other metrics proposed in recent literature.

Fix $p_1, p_2 \in [0,1]$ with $p_1 \geq p_2$. Then, we require that for a hash function $h \in \mathcal{H}, \Pr[h(x) = h(y)] \geq p_1 \forall\ x, y \in S$ with $d(x,y) \leq r$ for some fixed r. Additionally, for $x, y \in S$ with $d(x,y) > r, \Pr[h(x) = h(y)] \leq p_2$ for the same, fixed r. This intuitively allows for hashes to "collide" with probability p_1 when time series are r-similar. Further, the probability of incorrect collision is limited by p_2 (ideally close to 0) when two time series are more than r apart.

- Use hash that approximates time series distance [2, 3].
- Compose multiple kernels or hash families to obtain ranked "normality" metric to be used in classification or normality (anomaly) detection [4].
- Perform continuous hashing on rolling windows of execution across channels of multiple resources to classify activity in real-time.

References

[1] Daniel Lemire. Faster retrieval with a two-pass dynamic-time-warping lower bound. *Pattern Recogn.*, 42(9):2169–2180, September 2009.

[2] H. Hachiya and M. Matsugu. Nsh: Normality sensitive hashing for anomaly detection. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 795–802, Dec 2013.

[3] D. C. Kale, D. Gong, Z. Che, Y. Liu, G. Medioni, R. Wetzel, and P. Ross. An examination of multivariate time series hashing with applications to health care. In *2014 IEEE International Conference on Data Mining*, pages 260–269, Dec 2014.

[4] Y. B. Kim, E. Hemberg, and U. M. O'Reilly. Stratified locality-sensitive hashing for accelerated physiological time series retrieval. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2479–2483, Aug 2016.