Secure Binary Embeddings for Privacy Preserving Nearest Neighbors

- contributions of paper
  - a scheme for privacy preserving nearest neighbor search based on a secure stable embedding using quantized random projections
  - show how to use this scheme by presenting protocols for clustering and authentication applications

- quantization process
  - $x \in \mathbb{R}^K$, $y_m = \langle x, a_m \rangle + w_m$, $q_m = Q \left( \frac{y_m}{\Delta_m} \right)$
    - $x$: $K$-dimensional signal
    - $m = 1, \ldots, M$: measurement index
    - $y_m$: unquantized measurements
    - $a_m$: measurement vectors
    - $w_m$: additive dither
    - $\Delta_m$: quantization precision parameters
    - $Q(\cdot)$: quantizer

- universal quantization
  - scalar quantizer with non-contiguous quantization regions
  - determine an upper bound for the probability that there exist two signals $x$ and $x'$ with distance greater than $d$ that quantize to the same quantization vector given the number of measurements $M$
    - $P(q = q') = P(x, x' \text{ consistent}|d) = \frac{1}{2} + \sum_{i=0}^{+\infty} \frac{e^{-\left(\frac{\pi(2i+1)\sigma d}{\sqrt{2} \Delta}\right)^2}}{(\pi(i + 1/2))^2}$
    - $P(q = q') = P(x, x' \text{ consistent}|d) \leq \frac{1}{2} + \frac{1}{2} e^{-\left(\frac{\pi \sigma d\Delta}{\sqrt{2} \Delta}\right)}$
    - $d = \|x - x'\|_2$, $q = Q \left( \frac{\langle x, a \rangle + w}{\Delta} \right)$, $q' = Q \left( \frac{\langle x', a \rangle + w}{\Delta} \right)$
      - $a$: drawn from a normal distribution mean 0 and variance $\sigma^2$
      - $w$: uniformly distributed in $[0, \Delta]$
secure binary embeddings

- the quantization process used as an embedding has similar properties to Locality Sensitive Hashing (LSH)
- information-theoretic security:

\[
I(q_i; q'_i|d) = \sum_{q_i, q'_i \in \{0, 1\}} P(q_i, q'_i|d) \log \frac{P(q_i, q'_i|d)}{P(q_i|d)P(q'_i|d)}
\]

\[
I(q; q'|d) \leq 10Me^{-\left(\frac{\pi \sigma d}{\Delta}\right)^2}
\]

- stable embedding:

provide a relationship between the distance of the signals and the distance of their quantized measurements

binary space: \(\{0, 1\}^M \rightarrow \) Hamming distance: 

\[
d_H(q, q') = \frac{1}{M} \sum_m (q_m \oplus q'_m)
\]

with probability 

\[
1 - 2e^{2\log L - 2d^2M}, 1 - P_{c|d} - t \leq d_H(q, q') \leq 1 - P_{c|d} + t
\]

- \(L\): number of points to be embedded securely
- \(P_{c|d}\): shorthand for \(P(x, x' \text{ consistent}|d)\)
- \(t\): control variable
• applications
  – privacy preserving clustering with a star topology
  – authentication using symmetric keys
  – privacy preserving clustering with two parties
Information Retrieval Methods for Automatic Speech Recognition

- introduction
  - use n-gram features extracted from phonetic recognition, multi-phone recognition and word recognition
  - using IR for mapping acoustic features to word sequences provides a more flexible pronunciation model
  - the vector space model used in IR has no sequencing constraints, which leads to robustness against disfluencies and noise

- proposed speech recognition method

  ![Speech recognition using IR diagram](image)

  Figure 3: Speech recognition using IR.

  - recognition task
    * map the audio into a sequence of acoustic units
    * increasing the size of acoustic units (phone → multi-phone → word) leads to a decrease in the phonetic error rate, although at the cost of more complex models and the remaining errors are more difficult to correct
    * decreasing the size of acoustic units leads to an increase in the phonetic error rate, but the IR engine may have a good chance to recover from errors as long as enough phones are correctly recognized

  - feature extraction task
    * use the acoustic units to produce features useful for the task
    * bigram unit features:
      the complete set of bigrams is not large
      the bigrams contain more sequencing information than the unigrams
      they are more robust to recognition errors than longer units
    * maximum mutual information n-gram unit features:
      computed for units where a sufficient amount of data is available
IR scoring task

* vector space model based scoring:
  vectors consist of weights calculated from the training data
  \[
  \cos(v_q, v_d) = \frac{\sum_k v_{qk} v_{dk}}{|v_q| |v_d|}, \quad v_q: \text{testing query}, \ v_d: \text{training document}
  \]
  TF-IDF: \( v_{jk} = \frac{f_{jk}}{m_j} \left( 1 + \log_2 \left( \frac{n}{n_k} \right) \right), \ j = q, d \)
  document \( \hat{i} = \arg \max_i \cos(v_q, r_i) \)
  discriminative training: use minimal classification error criterion

* language model based scoring:
  document \( \hat{d} = \arg \max_d P(d|q) = \arg \max_d P(q|d)P(d) \)