Mathematical Pseudocode – Anomaly-Based Duplicate Detection: A Probabilistic Approach

Andreas Obermeier [0000-0002-0048-8208]

University of Ulm, 89069 Ulm, Germany
andreas.obermeier@uni-ulm.de

1 Interval Method

algorithm interval-method **is**

input: set of feature vectors $\{\zeta_{i,j} \in \mathbb{R}^f\}$ in the dataset to be analyzed
set of feature vectors $\{\hat{\zeta}_{i,j} \in \mathbb{R}^f\}$ in duplicate-free training data
multidimensional interval $I_h \subset \mathbb{R}^f$
number of samples N

output: duplicate probability $P\left(D \middle| \zeta_{i,j} \in I_h\right)$ for pairs of records in I_h

$m \leftarrow \left|\{\zeta_{i,j} \in \mathbb{R}^f\}\right|$ // sample size = size of the dataset to be analyzed
$\varrho_h \leftarrow \left|\left\{\zeta_{i,j} \middle| \zeta_{i,j} \in I_h\right\}\right|$ // count of feature vectors in I_h

for n=1 **to** N:
  $\{\hat{\zeta}^s_{i,j}\} \leftarrow$ sample of m feature vectors from $\{\hat{\zeta}_{i,j} \in \mathbb{R}^f\}$
  $\hat{\varrho}_{h,n} \leftarrow \left|\left\{\hat{\zeta}^s_{i,j} \middle| \hat{\zeta}^s_{i,j} \in I_h\right\}\right|$ // count of feature vectors in I_h in sample n

for k=1 **to** m:
  $\hat{P}(\hat{\varrho}_h = k) \leftarrow \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}_{\{\hat{\varrho}_{h,n}=k\}}$ // probability of k counts in training data

$\hat{E}(\hat{\varrho}_h) \leftarrow \sum_{k=1}^{m} k \cdot \hat{P}(\hat{\varrho}_h = k)$ // expected value in duplicate-free data

if $\varrho_h > \hat{E}(\hat{\varrho}_h)$: // test for anomaly
  $P\left(D \middle| \zeta_{i,j} \in I_h\right) \leftarrow \frac{\sum_{\varrho_h^D=0}^{\varrho_h} \hat{P}(\hat{\varrho}_h = \varrho_h - \varrho_h^D) \cdot \varrho_h^D}{\varrho_h}$ // estimated duplicate probability
  return $P\left(D \middle| \zeta_{i,j} \in I_h\right)$

else:
  return 0 // no anomaly, duplicate probability zero

2 Kernel Density Estimation Method

```
algorithm kde-method is

  input:  set of feature vectors {ζ_l ∈ ℝ^f} in the dataset to be analyzed with
            (1 ≤ l ≤ datset size) (for example, ζ_1 = ζ_{1,2})
          feature vector ζ_{i,j} ∈ {ζ_l} of pair of records in question
          set of feature vectors {ζ̂_l ∈ ℝ^f} in duplicate-free training data
          kernel function K for kernel density estimation
          number of samples N

  output: duplicate probability P(D|ζ_{i,j}) for a specific pair of records with
          feature vector ζ_{i,j}


  function KDE(x, {x_i}):       // function for Kernel Density Estimation
    L ← |{x_i}|
    return (1/L) ∑_{n=1}^{L} K(x − x_i)
  end function


  m ← |{ζ_{i,j} ∈ ℝ^f}|          // sample size = size of the dataset to be analyzed
  ϱ_{i,j} ← KDE(ζ_{i,j}, {ζ_l})     // estimated density at ζ_{i,j} in the dataset to be analyzed


  for n=1 to N:
    {ζ̂_l^s} ← sample of m feature vectors from {ζ̂_l}
    ϱ̂_{i,j}^{(n)} ← KDE(ζ_{i,j}, {ζ̂_l^s})           // estimated density at ζ_{i,j} in sample n


  {ϱ̂_{i,j}^{(n)}} ← {ϱ̂_{i,j}^{(1)}, ϱ̂_{i,j}^{(2)}, …, ϱ̂_{i,j}^{(N)}}     // set of sampled densities at ζ_{i,j}


  Ê(ϱ̂_{i,j}) ← ∫_{−∞}^{∞} ϱ̂_{i,j} · KDE(ϱ̂_{i,j}, {ϱ̂_{i,j}^{(n)}}) dϱ̂_{i,j}    // expected value in duplicate-free data


  if ϱ_{i,j} > Ê(ϱ̂_{i,j}):                        // test for anomaly
    P(D|ζ_{i,j}) ← [∫_0^{ϱ_{i,j}} KDE(ϱ_{i,j} − ϱ_{i,j}^D, {ϱ̂_{i,j}^{(n)}}) · ϱ_{i,j}^D dϱ_{i,j}^D] / ϱ_{i,j}    // estimated duplicate probability
    return P(D|ζ_{i,j})

  else:
    return 0                                      // no anomaly, duplicate probability zero
```