Improvement of Reliability Score for Autocoding and its Implementation in R

uRos2019, 21 May 2019, Bucharest

Yukako Toko¹, Shinya Iijima¹, Mika Sato-Ilic¹,²
¹National Statistics Center, Japan
²University of Tsukuba, Japan
Overview - Background

Uncertainty of Training Data
- Semantic problem
- Interpretation problem
- Insufficiently detailed input information

Development of Overlapping Classifier
One Feature into Multiple Classes
Utilized the idea of Fuzzy Partition Entropy

Conventional Classifier
One Feature into One Class

Problem

Uncertainty from data
Probability Measure
Uncertainty from latent classification structure in data
Fuzzy Measure

Reliability Score
Considering Uncertainties from Both Measures
Utilize Difference of Measures for Uncertainties

Purpose of This Study

Overlapping Classifier based on Reliability Score

Improvement of the reliability score

Consideration of Generalized Reliability Score

→ Apply T-norm in Statistical metric space

Consideration of Frequency of Each Feature in training dataset

→ Inclusion of the frequency of each feature to the Reliability Score
To address the unrealistic restriction: one feature is classified to a single class
-> proposed an algorithm that allows the assignment of one feature
is classified to multiple classes
Method – Overlapping classifier

Step 1: Calculate the probability of $j$-th feature ($j=1,…,J$) to a class $k$ ($k=1,…,K$) as

$$p_{j,k} = \frac{n_{j,k}}{n_j}, \quad n_j = \sum_{k=1}^{K} n_{j,k}$$

$n_{j,k}$ : Number of text descriptions in a class $k$ with $j$-th feature in the training dataset
Method – Overlapping classifier

Step 2: Determine at most $\tilde{K}$ ($\tilde{K} < K$) promising candidate classes for each feature based on $\tilde{p}_{jk}$

1. Arrange $\{p_{j1}, \ldots, p_{jK}\}$ in descending order and create $\{\tilde{p}_{j1}, \ldots, \tilde{p}_{jK}\}$, such as $\tilde{p}_{j1} \geq \ldots \geq \tilde{p}_{jK}, j = 1, \ldots, J$

2. Create $\{\tilde{p}_{j1}, \ldots, \tilde{p}_{j\tilde{K}}\}$, $\tilde{K}_j \leq \tilde{K} \leq K$

Note: When there are same values in $\{\tilde{p}_{j1}, \ldots, \tilde{p}_{jK}\}$, then we select as many as possible different $\tilde{K}_j$ classes for each feature $j$
Method – Overlapping classifier

Step 3 : Calculate the **Reliability Score** $\tilde{p}_{jk}$

$$
\tilde{p}_{jk} = \tilde{p}_{jk} \left( 1 + \sum_{m=1}^{\tilde{K}_j} \tilde{p}_{jm} \log_{K} \tilde{p}_{jm} \right), \quad j = 1, \ldots, J, \quad k = 1, \ldots, \tilde{K}_j
$$

When the number of target text descriptions is $T$, and each text description includes $h_l$ ($l = 1, \ldots, T$) features, corresponding $\tilde{p}_{jk}$ for $l$-th text description can be represented as

$$
\tilde{p}_{j_lk}, \quad j_l = 1, \ldots, h_l, \quad k = 1, \ldots, \tilde{K}_{j_l}, \quad l = 1, \ldots, T
$$

Reliability score of $j$-th feature included in $l$-th text description to a class $k$

Step 4 : Determine top $L$ ($L \in \{1, \ldots, \sum_{j_l=1}^{h_l} \tilde{K}_{j_l}\}$) candidate classes
Method – Overlapping classifier

Degree of Reliability

\[\tilde{p}_{jk} : \text{Reliability Score of } j\text{-th feature to a class } k\]

\[\tilde{p}_{jk} = \tilde{p}_{jk} \left(1 + \sum_{m=1}^{\tilde{K}_j} \tilde{p}_{jm} \log_{\tilde{K}} \tilde{p}_{jm}\right)\]

- Probability
  - Probability of feature \( j \) to class \( k \)
- Fuzzy
  - Classification status of feature \( j \) over the \( \tilde{K}_j \) classes
  - Transformation from \( \tilde{p}_{jk} \) to classification status of feature \( j \)

Explanation of the uncertainty of the training data

Utilization of the deference of measurements of uncertainty
Method – different fuzzy measurement

Apply another fuzzy measurement for reliability score

Partition coefficient for each feature $j$

$$PC_j = \sum_{k=1}^{K} \tilde{p}_{jk}^2, \quad j = 1, ..., J$$

(Classification status of feature $j$ over the $K$ classes)

Another degree of Reliability

$$\bar{p}_{jk} = \tilde{p}_{jk} \sum_{k=1}^{K} \tilde{p}_{jk}^2$$

$$\bar{p}_{jk} = \tilde{p}_{jk} \left( 1 + \sum_{m=1}^{\tilde{K}_j} \tilde{p}_{jm} \log K \tilde{p}_{jm} \right)$$

Partition coefficient

Partition entropy
Method – Generalized Reliability Score

\[ \tilde{p}_{jk} = \sum_{k=1}^{K} \tilde{p}_{jk}^2 \]

Partition coefficient

\[ \tilde{p}_{jk} = \tilde{p}_{jk} \left( 1 + \sum_{m=1}^{\tilde{K}_j} \tilde{p}_{jm} \log_K \tilde{p}_{jm} \right) \]

Partition entropy

\[ \tilde{p}_{jk} = T \left( \tilde{p}_{jk} , \sum_{k=1}^{K} \tilde{p}_{jk}^2 \right) \]

\[ \tilde{p}_{jk} = T \left( \tilde{p}_{jk} , 1 + \sum_{m=1}^{\tilde{K}_j} \tilde{p}_{jm} \log_K \tilde{p}_{jm} \right) \]

Generalization

\( T(a, b) \): \textbf{T-norm} between a and b
Method – T-norm (Triangular norms)

\[ T : [0,1] \times [0,1] \rightarrow [0,1] \]
\[ \forall a, b, c, d \in [0,1] \]

(1) \[ 0 \leq T(a, b) \leq 1, \]
\[ T(a, 0) = T(0, b) = 0 \] (Boundary conditions)
\[ T(a, 1) = T(1, a) = a \]

(2) \[ a \leq c, b \leq d \implies T(a, b) \leq T(c, d) \] (Monotonicity)

(3) \[ T(a, b) = T(b, a) \] (Symmetry)

(4) \[ T(T(a, b), c) = T(a, T(b, c)) \] (Associativity)

(K. Menger, 1942)
Method – Statistical metric space

\[ F_{pq}(x) \equiv \Pr\{d_{pq} < x\} \]
\[ \forall p, q, r \in S \]

\[ d_{pp} = 0 \quad \leftrightarrow \quad F_{pp}(x) = 1, \text{ for all } x > 0 \]

\[ d_{pq} > 0 \ (p \neq q) \quad \leftrightarrow \quad F_{pq}(x) < 1, \ (p \neq q) \text{ for some } x > 0 \]

\[ d_{pq} = d_{qp} \quad \leftrightarrow \quad F_{pq} = F_{qp} \]

\[ d_{pr} \leq d_{pq} + d_{qr} \quad \leftrightarrow \quad F_{pr}(x + y) \geq T(F_{qp}(x), F_{qr}(y)) \]
Method – Examples of T-norm

<table>
<thead>
<tr>
<th>t-norm</th>
<th>( t(x, y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algebraic Prod.</td>
<td>( xy )</td>
</tr>
<tr>
<td>Hamacher Prod.</td>
<td>( \frac{xy}{p + (1 - p)(x + y - xy)} )</td>
</tr>
<tr>
<td>(( p \geq 0 ))</td>
<td></td>
</tr>
<tr>
<td>Sin based t-norm</td>
<td>( \frac{2}{\pi} \sin^{-1} \left[ \left( \sin \frac{\pi}{2} x + \sin \frac{\pi}{2} y - 1 \right) \vee 0 \right] )</td>
</tr>
<tr>
<td>Dombi Prod.</td>
<td>( \frac{1}{1 + \sqrt[1 - p]{ \left( \frac{1 - x}{x} \right)^p + \left( \frac{1 - y}{y} \right)^p } } )</td>
</tr>
<tr>
<td>(( p &gt; 0 ))</td>
<td></td>
</tr>
</tbody>
</table>
Method – Utilization of T-norm for Reliability Score

Algebraic Prod.

\[ \bar{p}_{jk} = \tilde{p}_{jk} * \sum_{k=1}^{K} \tilde{p}_{jk}^2 \]

Hamacher Prod.  
\( p \geq 0 \)

\[ \bar{p}_{jk} = \frac{\tilde{p}_{jk} \sum_{k=1}^{K} \tilde{p}_{jk}^2}{p + (1 - p)(\tilde{p}_{jk} + \sum_{k=1}^{K} \tilde{p}_{jk}^2 - \tilde{p}_{jk} \sum_{k=1}^{K} \tilde{p}_{jk}^2)} \]

Sin based t-norm

\[ \bar{p}_{jk} = \frac{2}{\pi} \sin^{-1} \left[ \sin \left( \frac{\pi}{2} \tilde{p}_{jk} + \sin \left( \frac{\pi}{2} \sum_{k=1}^{K} \tilde{p}_{jk}^2 - 1 \right) \right) \lor 0 \right] \]

Dombi Prod.  
\( p > 0 \)

\[ \bar{p}_{jk} = \frac{1}{1 + p \sqrt{\left( \frac{1 - \tilde{p}_{jk}}{\tilde{p}_{jk}} \right)^p + \left( \frac{1 - \sum_{k=1}^{K} \tilde{p}_{jk}^2}{\sum_{k=1}^{K} \tilde{p}_{jk}^2} \right)^p}} \]
Method – **Improved Reliability Score**

Utilize **T-norm** and **Sigmoid function**

\[
\tilde{p}_{jk} \equiv \text{tanh}(n_j) \ast \tilde{p}_{jk}
\]

\[
\tilde{p}_{jk} \equiv \frac{n_j}{\sqrt{1 + n_j^2}} \ast \tilde{p}_{jk}
\]

\[
\tilde{p}_{jk} = T\left(\tilde{p}_{jk}, \sum_{k=1}^{K} \tilde{p}_{jk}^2\right)
\]

\[
\tilde{p}_{jk} = T\left(\tilde{p}_{jk}, 1 + \sum_{m=1}^{K_j} \tilde{p}_{jm} \log_{K} \tilde{p}_{jm}\right)
\]
## Results

**Data: Family Income and Expenditure survey, Japan**

We used data answered via online survey system

**Data size:** approx. 400,000 instances

approx. 350,000 for Training

40,000 for Evaluation

### Results

<table>
<thead>
<tr>
<th>Number of total instances</th>
<th>Number of matched instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>---------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>1st candidate</td>
<td>35,044</td>
</tr>
<tr>
<td>2nd candidate</td>
<td>1,649</td>
</tr>
<tr>
<td>3rd candidate</td>
<td>536</td>
</tr>
<tr>
<td>4th candidate</td>
<td>283</td>
</tr>
<tr>
<td>5th candidate</td>
<td>189</td>
</tr>
<tr>
<td>Total</td>
<td>37,701</td>
</tr>
</tbody>
</table>

Hamacher Prod. $\frac{xy}{p + (1 - p)(x + y - xy)}$

($p = 0.99(PE), 0.7(PC)$)

Sigmoid func.

(a): $n_j/\sqrt{1 + n_j^2}$, (b): $\tanh n_j$
## Results

### Data: Family Income and Expenditure survey, Japan

Only foodstuff and dining-out data were used. We assigned 11 classification codes for this experiment.

**Data size**: 11,000 instances

- **10,000 for Training**
- **1,000 for Evaluation**

### Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1st candidate</td>
<td>854</td>
<td>854</td>
<td>854</td>
<td>854</td>
</tr>
<tr>
<td>2nd candidate</td>
<td>58</td>
<td>55</td>
<td>58</td>
<td>56</td>
</tr>
<tr>
<td>3rd candidate</td>
<td>20</td>
<td>26</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>932</strong></td>
<td><strong>935</strong></td>
<td><strong>932</strong></td>
<td><strong>933</strong></td>
</tr>
</tbody>
</table>

Data size: 11,000 instances

10,000 for Training

1,000 for Evaluation
We implemented this technique in R and the R package is under development.
Propose Generalized Reliability Score to Improve Handling Ability of the Reliability of Each Data to Each Code

Utilize T-norm in Statistical Metric Space to the Reliability Score

→ Generalize the Reliability Score

Inclusion of Frequency of Each Feature to the Reliability Score

Numerical examples show better performance

→ Improved Classification Accuracy

We implemented this technique in R and the R package is under development.
Reference


Thank you!
ytoko@nstac.go.jp