Improvement of Reliability Score for Autocoding and its Implementation in R

uRos2019, 21 May 2019, Bucharest

Yukako Toko[1], Shinya Iijima[1], Mika Sato-Ilic[1,2]
[1]National Statistics Center, Japan
[2] University of Tsukuba, Japan

Overview - Background

Conventional Classifier

One Feature
into One Class

Problem

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**

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

(Y. Toko, S. Iijima, M. Sato-Ilic, 2018)

Purpose of This Study

Overlapping Classifier
based on Reliability Score

(Y. Toko, S. Iijima, M. Sato-Ilic, 2018)

Improvement of the reliability score

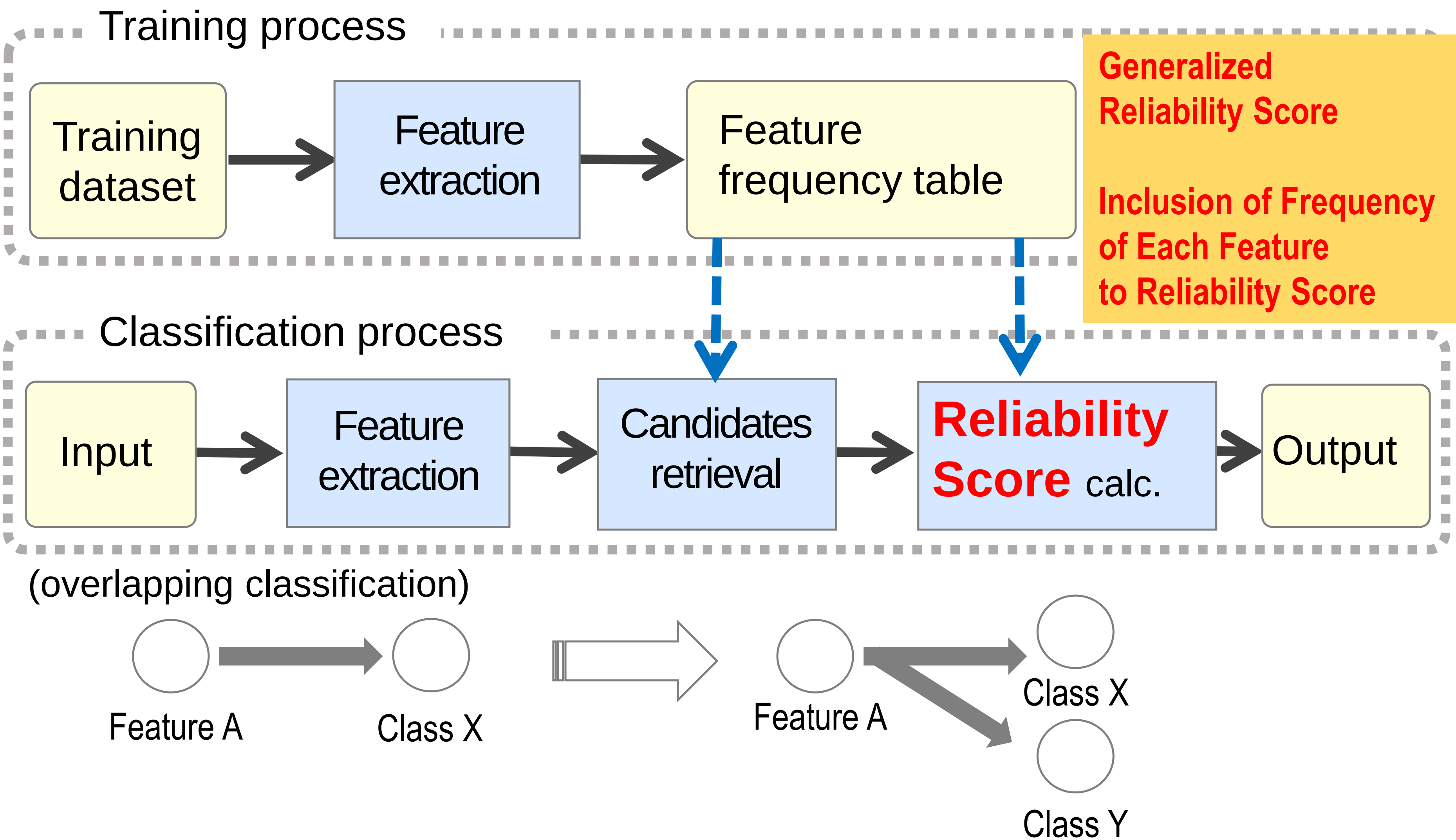
Consideration of **Generalized Reliability Score**

→ 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**

Overview - System Structure



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 ($j=1,\ldots,J$) to a class k ($k=1,\ldots,K$) as

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

n_{jk} : Number of text descriptions in a class k with j-th feature in the training dataset

Method – Overlapping classifier

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

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

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

Note : When there are same values in $\{\tilde{p}_{j1}, \cdots, \tilde{p}_{jK}\}$,
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** $\bar{p}_{jk}$

$$\boxed{\bar{p}_{jk} = \tilde{\tilde{p}}_{jk}\left(1 + \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm} \log_K \tilde{\tilde{p}}_{jm}\right), \quad j = 1, \cdots, J, \qquad k = 1, \cdots, \tilde{K}_j}$$

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

$$\bar{p}_{j_l k}, \qquad j_l = 1, \cdots, h_l, \qquad k = 1, \cdots, \tilde{K}_{j_l}, \qquad l = 1, \cdots, T$$

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

Step 4 : Determine top *L* ($L \in \{1, \dots, \sum_{j_l=1}^{h_l} \tilde{K}_{j_l}\}$) candidate classes

Method – Overlapping classifier

Degree of Reliability

$\bar{p}_{jk}$: Reliability Score of *j* -th feature to a class *k*

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

Explanation of the uncertainty of the training data

Utilization of the deference of measurements of uncertainty

Probability

Probability of feature *j* to class *k*

Fuzzy

Classification status of feature *j* over the $\tilde{K}_j$ classes

Transformation from $\tilde{\bar{p}}_{j\,k}$ to classification status of feature *j*

Method – different fuzzy measurement

Apply another fuzzy measurement for reliability score

Partition coefficient for each feature *j*

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

(Y. Toko, K. Wada, S. Iijima, M. Sato-Ilic, 2018)

Classification status of feature *j* over the K classes

Another degree of Reliability

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

Partition coefficient

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

Partition entropy

Method – **Generalized Reliability Score**

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

Partition coefficient

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

Partition entropy

Generalization

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

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

$\mathrm{T}(a\, , b)$: **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) $0 \leq T(a,b) \leq 1,$

$T(a,0) = T(0,b) = 0$

$T(a,1) = T(1,a) = a$

(Boundary conditions)

(2) $a \leq c, b \leq d \Rightarrow T(a,b) \leq T(c,d)$ (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

$$F_{pq}(x) \equiv \Pr\{d_{pq} < x\}$$

$$\forall p, q, r \in S$$

$d_{pp} = 0$	$\leftrightarrow$	$F_{pp}(x) = 1, \text{for all } x > 0$
$d_{pq} > 0 \ (p \neq q)$	$\leftrightarrow$	$F_{pq}(x) < 1, (p \neq q) \text{ for some } x > 0$
$d_{pq} = d_{qp}$	$\leftrightarrow$	$F_{pq} = F_{qp}$
$d_{pr} \leq d_{pq} + d_{qr}$	$\leftrightarrow$	$F_{pr}(x + y) \geq T(F_{qp}(x), F_{qr}(y))$

Method – Examples of T-norm

t-norm	$t(x,y)$
Algebraic Prod.	xy
Hamacher Prod. $(p \geq 0)$	$\dfrac{xy}{p+(1-p)(x+y-xy)}$
Sin based t-norm	$\dfrac{2}{\pi} sin^{-1}\left[\left(\sin\dfrac{\pi}{2}x+\sin\dfrac{\pi}{2}y-1\right)\vee 0\right]$
Dombi Prod. $(p > 0)$	$\dfrac{1}{1+\sqrt[p]{\left(\dfrac{1-x}{x}\right)^p+\left(\dfrac{1-y}{y}\right)^p}}$

Method – Utilization of T-norm for Reliability Score

Algebraic Prod.

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

Hamacher Prod. $(\, p \geq 0 \,)$

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

Sin based t-norm

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

Dombi Prod. $(\, p > 0 \,)$

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

Method – Improved Reliability Score

Utilize T-norm and Sigmoid function

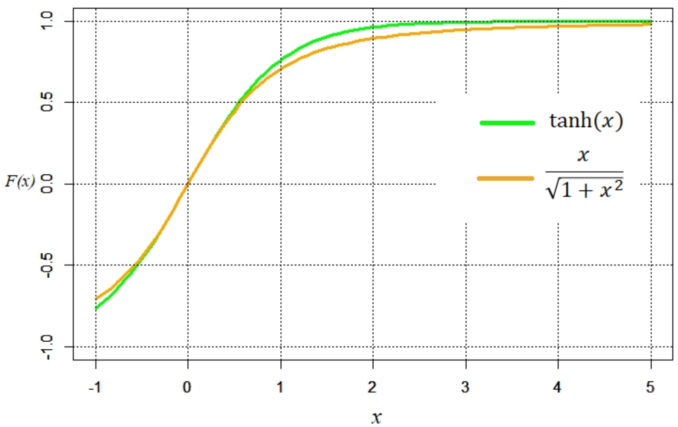
$$\bar{\bar{p}}_{jk} \equiv \tanh(n_j) * \bar{p}_{jk}$$

$$\bar{\bar{p}}_{jk} \equiv \frac{n_j}{\sqrt{1 + {n_j}^2}} * \bar{p}_{jk}$$

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

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

T-norm



Sigmoid function

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

	Number of total instances	Number of matched instances					
		Partition Entropy (PE) +Algebraic Prod.	Partition Coefficient (PC) +Algebraic Prod.	PE + Hamacher Prod. + Sigmoid func. (a)	PE + Hamacher Prod. + Sigmoid func. (b)	PC + Hamacher Prod. + Sigmoid func. (a)	PC + Hamacher Prod. + Sigmoid func. (b)
1st candidate	40,000	35,044	35,051	35,064	35,100	35,119	35,134
2nd candidate		1,649	1,682	1,618	1,589	1,614	1,595
3rd candidate		536	540	551	541	539	539
4th candidate		283	293	277	283	291	293
5th candidate		189	179	189	187	185	188
Total		37,701	37,745	37,699	37,700	37,748	37,749

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

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

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

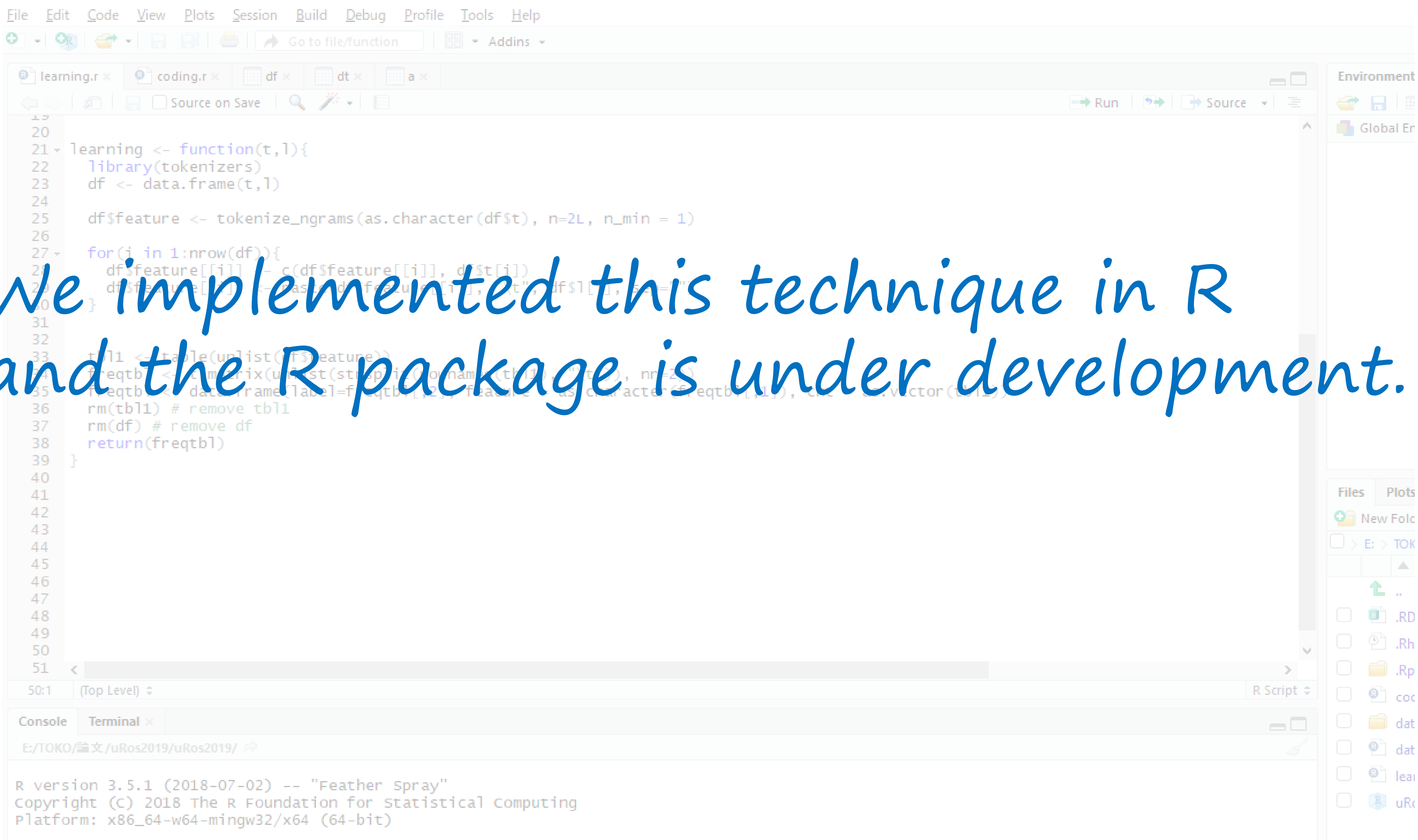
Data size : 11,000 instances

10,000 for Training
1,000 for Evaluation

Results

	Number of total instances	Number of matched instances			
		T-norm			
		Algebraic Prod.	Sin-based T-norm	Hamacher Prod.	Dombi Prod.
1st candidate	1,000	854	854	854	854
2nd candidate		58	55	58	56
3rd candidate		20	26	20	23
Total		932	935	932	933

Implementation in R



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

Summary

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

[1] B. Schweizer, A. Sklar, Probabilistic Metric Spaces, Dover Publications, New York (2005).

[2] J. C. Bezdek, J. Keller, R. Krisnapuram, N.R. Pal, Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, Kluwer Academic Publishers, New York (1999).

[3] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York (1981).

[4] K. Menger, Statistical Metrics, Proc. Nat. Acad. of Sci,. U.S.A., 28 (1942), 535-537.

[5] M. Mizumoto, Pictorial Representaition of Fuzzy connectives, Part I: Cases of t-Norms, t-Conorms and Averaging Operators, Fuzzy Sets and Systems, 31 (1989), 217-242.

[6] Y. Toko, S. Iijima, M. Sato-Ilic, Overlapping classification for autocoding system, Journal of Romanian Statistical Review, Vol. 4, (2018), 58-73.

[7] Y. Toko, K. Wada, S. Iijima, M. Sato-Ilic, Supervised multiclass classifier for autocoding based on partition coefficient, Intelligent Decision Technologies 2018, Smart Innovation, Systems and Technologies, Springer, Switzerland, Vol. 97, (2018), 54-64.

Thank you !

ytoko@nstac.go.jp