Generalization for Improvement of the Reliability Score for Autocoding

Yukako Toko (ytoko@nstac.go.jp)
National Statistics Center, Japan

Shinya Iijima (siijima@nstac.go.jp)
National Statistics Center, Japan

Mika Sato-Ilic (msato@nstac.go.jp)
National Statistics Center, Japan / University of Tsukuba

ABSTRACT

We developed the supervised multiclass classifier for autocoding in our previous study. The classifier assigns corresponding classification codes based on reliability scores. The purpose of this study is the generalization of the reliability score for more accurate classification.

The previously defined reliability score considering both the uncertainty from data (probability measure) and the uncertainty from the latent classification structure in data (fuzzy measure) gives our method a better accuracy of the result. However, this reliability score still has problems to be addressed. The first problem is that the reliability score has not been generalized. When we consider applying our classifier in a practical situation, we must consider variability of the observed data. Therefore, there is a necessity to generalize the reliability score. The second problem is that the reliability score does not consider the frequency for each object (or feature) which means a sum of frequencies over the codes for each object (or feature) in the training dataset. This problem could cause infrequent objects (or features) to have significant influence leading to the classifier sometimes incorrectly classifying data.

To overcome these problems, we propose a generalized reliability score by using the idea of the T-norm in statistical metric space and considering the frequency of each object (or feature) over codes in data. In addition, we investigate the robustness of the proposed classifier based on the generalized reliability score to show the guarantee of the classification accuracy based on the classifier.

The proposed algorithm is implemented in R to improve its versatility.

Keywords: *Coding, Machine learning, Overlapping classification, Generalization of reliability score*

JEL Classification:*C38*

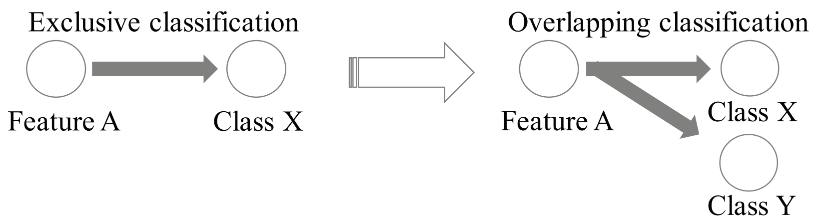
1. INTRODUCTION

In official statistics, text response fields such as fields for occupation, industry, and the type of household income and expenditure are sometimes found on survey forms. Those respondents' text descriptions are usually translated into corresponding classification codes for efficient data processing. Originally, coding tasks are performed manually, whereas the studies of automated coding have made progress with the improvement of computer technology in recent years. For example, Hacking and Willenborg (2012) introduced coding methods, including autocoding. Gweon et al. (2017) illustrated methods for automated occupation coding based on statistical learning.

We also developed a supervised multiclass classifier for autocoding in the coding task for *the Family Income and Expenditure Survey* in Japan. Initially, the classifier we developed was based on a simple machine learning technique (Toko et al., 2017; Shimono et al., 2018). It selects the most promising classification code based on the probability provided from training data. However, the classifier incorrectly classifies some text descriptions with ambiguous information because of the semantic problem, interpretation problem, and insufficiently detailed input information. As the main reason for these problems is the unrealistic restriction that one object (or feature) is classified to a single class, we developed a new classifier (Toko et al., 2018a; Toko et al., 2018b) that allows the assignment of one object (or feature) to be classified into multiple classes with a calculation of the reliability scores based on partition entropy or partition coefficient (Bezdek, 1981; Bezdek et al., 1999). Fig.1 shows the basic idea of the overlapping classifier. This overlapping classifier assigns classification codes considering both the uncertainty from training data (probability measure) and the uncertainty from latent classification structure in data (fuzzy measure).

The basic idea of the overlapping classifier

Fig. 1



Although we improved the classification accuracy of the classifier in our previous studies, the classifier still has tasks to be addressed. One task is to consider the generalization of the reliability score and the guarantee of the robustness of the overlapping classification in the practical situation. In addition, the previously proposed reliability score does not consider the frequency of each object (or feature) over the codes in the training dataset; we cannot consider the confidence degree of existence of each object (or feature). This could cause an infrequent object (or feature) to have significant influence leading to the classifier sometimes incorrectly classifying data. To achieve these tasks, this study implements two approaches. First, we propose generalized reliability scores defined by using the idea of *T*-norm in statistical metric space (Menger, 1942; Mizumoto, 1989; Schweizer and Sklar, 2005) considering the frequency of each object (or feature) over codes in data as the confidence of each object (or feature). Second, we investigate the robustness of the overlapping classifier with the generalized reliability score. A classifier for the autocoding system in official statistics requires the stable code assignment, whereas the style of text description is not always stable even in the same survey as it depends on respondents.

Our classifier is developed in R for the following reasons: First, the abundant packages available in R allow for the efficient development of the classifier. We used packages such as “tokenizers” (Mullen et al., 2018), “dplyr” (Wickham et al., 2018) and “data.table” (Dowle and Srinivasan, 2019) for efficient processing. Second, when implemented in R, our classifier could be a better contribution to coding tasks in the field of official statistics as the number of R users is increasing in this field. An R package of our classifier is under development.

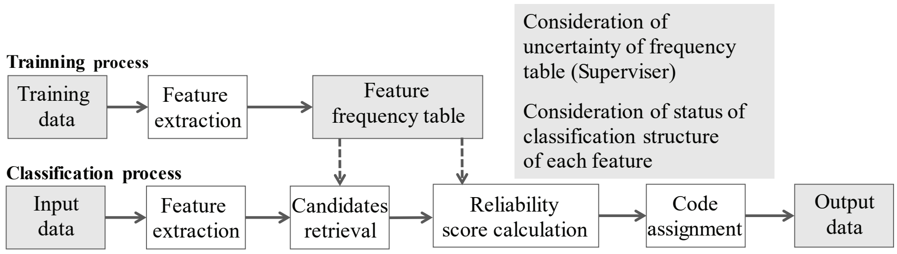
The rest of this paper is organized as follows: The previously developed overlapping classifier (Toko et al., 2018b) is described in section 2, the generalized reliability score is proposed in section 3. The numerical examples are illustrated in section 4. Conclusions and suggestions for future work are described in section 5.

2. PREVIOUSLY DEVELOPED OVERLAPPING CLASSIFIER

The overlapping classifier (Toko et al., 2018a; Toko et al., 2018b) allows the assignment of one object (or feature) into multiple codes, and it comprises training and classification processes, as shown in Fig. 2.

Basic structure of our classifier

Fig. 2



In the training process, objects (or features) are extracted, and a feature frequency table is created. We took word-level N-grams from the word sequences of text descriptions using "tokenizers" (Mullen et al., 2018) package after tokenizing each description using MeCab (Kudo et al., 2004), which is a well-known dictionary-attached morphological Japanese text analyzer. Here, unigrams (any word), bigrams (any sequence of two consecutive words), and entire sentences are considered as objects (or features). After feature extraction, the system tabulates all extracted objects (or features) based on their given classification codes into a feature frequency table.

In the classification process, the overlapping classifier performs feature extraction, the retrieval of candidate classification codes, and classification codes assignment. First, the classifier extracts objects (or features) of target text descriptions. Then, it retrieves the corresponding classification codes from the feature frequency table provided by using the extracted objects (or features). After that, the classifier calculates the probability of *j*-th object (or feature) assigned to a code k, under the assumption that random variable for *j*-th object (or feature) being 1, that is defined as

$$p_{jk} = \frac{n_{jk}}{n_j}, \quad n_j = \sum_{k=1}^{K} n_{jk}, \quad j = 1, \dots, J, \qquad k = 1, \dots, K, \tag{1}$$

where n_{jk} is the number of occurrence of statuses in which an object (or feature) j assigned to a code k in the training dataset. Then, we arrange $\{p_{j1}, \cdots, p_{jK}\}$ in descending order and create $\{\tilde{p}_{j1}, \cdots, \tilde{p}_{jK}\}$, such as $\tilde{p}_{j1} \geq \cdots \geq \tilde{p}_{jK}, j = 1, \cdots, J$. Next, we select at most $\tilde{K}$, $(\tilde{K} \leq K)$ promising candidate codes for each object (or feature) based on the values of $\tilde{p}_{jk}$. That is, we create $\{\tilde{\tilde{p}}_{j1}, \cdots, \tilde{\tilde{p}}_{j\tilde{K}_j}\}$, $\tilde{K}_j \leq \tilde{K} \leq K$. In the case when we cannot select different $\tilde{K}$ codes, that is the case when there are the same values in $\{\tilde{p}_{j1}, \cdots, \tilde{p}_{jK}\}$, then we select as many different $\tilde{K}_j$ codes as possible for each object (or feature) j. Then, we define the reliability score $\bar{p}_{jk}$ utilizing the partition entropy or partition coefficient (Bezdek, 1981; Bezdek et al., 1999) as follows

$$\bar{p}_{jk} = \tilde{\tilde{p}}_{jk}\left(1 + \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm} \log_K \tilde{\tilde{p}}_{jm}\right), \quad j = 1, \cdots, J, \quad k = 1, \cdots, \tilde{K}_j. \tag{2}$$

$$\bar{p}_{jk} = \tilde{\tilde{p}}_{jk} \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm}^2, \qquad j = 1, \dots, J, \qquad k = 1, \dots, \tilde{K}_j. \tag{3}$$

When the number of target text descriptions is L, and each text description includes h_l, $l = 1, \cdots, L$ objects (or features), corresponding $\bar{p}_{jk}$ shown in (2) for l-th text description can be represented as

$$\bar{p}_{j_l k}, \; j_l = 1, \cdots, h_l, \; k = 1, \cdots, \tilde{K}_{j_l}, \quad l = 1, \cdots, L, \tag{4}$$

which shows a reliability score of j-th object (or feature) included in l-th text description to a code k. The total number of the promising candidate codes for l-th text description is $\sum_{j_l=1}^{h_l} \tilde{K}_{j_l}$. Then, the classifier selects top $V \in \{1, \dots, \sum_{j_l=1}^{h_l} \tilde{K}_{j_l}\}$ codes for assignment of l-th text description based on the reliability score $\bar{p}_{j_l k}$ shown in (4).

3. GENERALIZED RELIABILITY SCORE

We have handled many kinds of data, and the observed data has become more complicated. This means that it might be difficult to apply a single defined aggregation operator that retrieves the probabilistic uncertainty and fuzzy uncertainty for defining the reliability score. Therefore, we extend the definition of the aggregation operator as T-norm including the previously defined aggregation operator as algebraic product, and generalize the reliability score by using the idea of T-norm, which is a binary operator in statistical metric space. T-norm satisfies the following four conditions:

- Boundary conditions

$$0 \leq T(a,b) \leq 1, \qquad T(a,0) = T(0,b) = 0, \qquad T(a,1) = T(1,a) = 1$$

- Monotonicity

$$a \leq c\,, b \leq d \;\rightarrow T(a,b) \leq T(c,d)$$

- Symmetry

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

- Associativity

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

where $\forall a, b, c, d \in [0,1]$.

As there are numerous possible choices for *T*-norms, once we generalized the reliability score, we can select an appropriate *T*-norm from various choices of *T* according to data. The following *T*-norms are employed for $\forall x, y \in [0,1]$ in this study:

- Algebraic product

$$T(x,y) = xy$$

- Sin-based *T*-norm

$$T(x,y) = \frac{2}{\pi}\sin^{-1}\left[\left(\sin\frac{\pi}{2}x + \sin\frac{\pi}{2}y - 1\right) \vee 0\right]$$

- Hamacher product

$$T(x,y) = \frac{xy}{p + (1-p)(x + y - xy)}, \quad p \geq 0$$

- Dombi product

$$T(x,y) = \frac{1}{1 + \sqrt[p]{\left(\frac{1-x}{x}\right)^p + \left(\frac{1-y}{y}\right)^p}}, \qquad p > 0$$

Then, the generalized reliability scores utilizing *T*-norm are defined as

$$\bar{\bar{p}}_{jk} = T\left(\tilde{\tilde{p}}_{jk}, 1 + \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm} \log_K \tilde{\tilde{p}}_{jm}\right), \quad j = 1, \cdots, J, \quad k = 1, \cdots, \tilde{K}_j. \tag{5}$$

$$\bar{\bar{p}}_{jk} = T\left(\tilde{\tilde{p}}_{jk}, \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm}^2\right), \quad j = 1, \ldots, J, \quad k = 1, \ldots, \tilde{K}_j. \tag{6}$$

Our original proposed reliability scores shown in (2) and (3) are also one of the generalized reliability scores, which is the case when we use the algebraic product as the *T*-norm in (5) and (6). This means that the originally proposed reliability scores are special cases in the proposed generalized reliability scores. In these generalized reliability scores, we consider the two values obtained by probability measure and fuzzy measure. That is, in (5) and (6), $\tilde{\tilde{p}}_{jk}$ shows the uncertainty from training data (probability measure) and $1 + \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm} \log_K \tilde{\tilde{p}}_{jm}$ or $\sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm}^2$ are the uncertainty from latent classification structure in data (fuzzy measure). These values of the uncertainty from the latent classification structure can show the classification status of each object (or feature); that is how each object (or feature) classified to the candidate codes (or classes). If an object (or feature) indicates all of the candidate codes evenly, then the values become smaller, otherwise, if an object (or feature) indicates clearly a candidate code, then the values become larger. Therefore, the reliability scores shown in (5) and (6) consider not only the uncertainty of an object (or feature) *j* to a code *k* which is obtained directly from the training data, but also the classification status of the object (or feature) *k* to all of the candidate codes. And the combining operator between the two values of two different uncertainties is generalized by using *T*-norm in the statistical metric space.

For the generalized reliability score, we consider the frequency of each object (or feature) to the reliability score by using following functions:

$$g(\theta) = \frac{\theta}{\sqrt{1+\theta^2}}. \tag{7}$$

$$g(\theta) = \tanh(\theta). \tag{8}$$

These functions satisfying the following conditions:

- $-\infty \leq \theta_1 \leq \infty \rightarrow -1 \leq g(\theta_1) \leq 1$
- $\theta_1 \leq \theta_2 \rightarrow g(\theta_1) \leq g(\theta_2)$
- $g(0) = 0$
- $\lim_{\theta_1 \to \infty} g(\theta_1) = 1$

where $\forall \theta_1, \theta_2 \in \mathbf{R}$. Then, we redefine the generalized reliability score considering the frequency of each object (or feature) as follows:

$$\bar{\bar{\bar{p}}}_{jk} = g(n_j) \times \bar{\bar{p}}_{jk}. \qquad (9)$$

4. INVESTIGATION OF ROBUSTNESS FOR PROPOSED CLASSIFIER

For investigating the robustness of the overlapping classifier with the generalized reliability scores, we perform the following procedures. First, we perform the normal overlapping classification. Then, we generate a normal random number δ_{jk} ($\delta_{jk} \sim N(0, \sigma^2)$) as a noise for each probability of *j*-th object (or feature) to a code k. After that, we add the noise δ_{jk} for each probability of *j*-th object (or feature) to a code k. Then, we again perform the codes assignment based on the noise added probability. Finally, we compare the difference of results between the normal overlapping classification and the overlapping classification based on the value of noise added probability of *j*-th object (or feature) to a code k.

(Step 1) Perform the normal class assignment using the overlapping classifier: extract objects (or features), calculate p_{jk} and $\bar{\bar{\bar{p}}}_{j_l k}$ shown in (1) and (9). Set $\bar{\bar{\bar{p}}}_{j_l k}^{(0)} = \bar{\bar{\bar{p}}}_{j_l k}$. The result of this class assignment is used as a benchmark.

(Step 2) Generate normal random numbers as

$$\Delta = (\delta_{jk}), \quad j = 1, \ldots, J, \quad k = 1, \ldots, K, \quad \delta_{jk} \sim N(0, \sigma^2). \qquad (10)$$

(Step 3) Calculate $\hat{p}_{jk} \equiv p_{jk} + \delta_{jk}$. Select at most $\tilde{K}$ ($\tilde{K} \leq K$) promising candidates for each object (or feature) based on calculated $\hat{p}_{jk}$.

(Step 4) Calculate reliability scores $\bar{\bar{\bar{p}}}_{j_l k}$ by using (9) based on $\hat{p}_{jk}$.

(Step 5) Determine top $V \in \{1, \ldots, \sum_{j_l=1}^{h_l} \tilde{K}_{j_l}\}$ classes based on the reliability scores.

(Step 6) Set different σ shown in (10) and repeat Step 2 to Step 5. Let M_{ni} be the number of text descriptions that match with i-th candidate class under n-th different σ, $n = 1, \dots, Q$ and let M_{1i} be the number of text descriptions that match with i-th candidate class under the use of $\bar{\bar{\bar{p}}}_{j_l k}^{(0)}$ in Step 1. Then, we introduce the difference of classification accuracy compared to the normal classification as

$$d_{ni} = \frac{M_{ni}}{M_{1i}}, n = 1, \dots, Q, i = 1, \dots, V. \tag{11}$$

5. NUMERICAL EXAMPLES

For the numerical example, we applied our overlapping classifier with generalized reliability scores shown in (9) to *the Family Income and Expenditure Survey* dataset. We used approximately 400 thousand instances for this experiment and randomly divided this dataset into 40 thousand instances for evaluation and the rest of them for training. Each instance contains an item name (a transacted item name that is related to income or expenditure) in Japanese and a manually assigned corresponding classification code (more than 550 classification codes available). For efficient classification, we separately performed the training and classification process for income and expenditure data.

Table 1 shows the result of the code assignment using each generalized reliability score. For the generalized reliability used partition coefficient shown in (6), we use the case when $\tilde{K} = K$, that is we use all of the codes. The results of the code assignment using generalized reliability scores were almost always better than the result of the code assignment using the original reliability score. In particular, the case when we use the partition coefficient for the fuzzy measurement, Hamacher product when $p = 0.7$ for T-norm, and the function g shown in (8), was the best total result. In addition, the numbers of matched instances at 1st candidate code are improved compared with the result of the code assignment using the original reliability score. Note that, "Sigmoid func. (a)" and "Sigmoid func. (b)" in tables below means $g(n_j) = n_j/\sqrt{1 + n_j^2}$ shown in (7), and $g(n_j) = \tanh(n_j)$ shown in (8) respectively.

Classification accuracy of each reliability score

Table 1

Condition of reliability score (R.S.)						Number of total instances	Number of matched instances					
							1st candidate	2nd candidate	3rd candidate	4th candidate	5th candidate	Total
Fuzzy measurement	Partition Entropy	T-norm	Algebraic prod.	$g(n_i)$	NA (Original)	40,000	35,044	1,649	536	283	189	37,701
					Sigmoid func. (a)		35,064	1,618	552	277	188	37,699
					Sigmoid func. (b)		35,100	1,589	541	283	187	37,700
			Hamacher prod.		Sigmoid func. (a)		35,051	1,682	540	293	179	37,745
					Sigmoid func. (b)		35,100	1,589	541	283	187	37,700
	Partition Coefficient		Algebraic prod.		Sigmoid func. (a)		35,118	1,617	540	292	178	37,745
					Sigmoid func. (b)		35,137	1,595	541	293	178	37,744
			Hamacher prod.		Sigmoid func. (a)		35,119	1,614	539	291	185	37,748
					Sigmoid func. (b)		35,134	1,595	539	293	188	37,749

Table 2 shows the classification accuracy of data that have the value of the generalized reliability score is 0.99 or more at 1st candidate. From Table 2, it seems that the generalized reliability scores have more consistency with the classification accuracy as they consider the frequency of each object (or feature) in training dataset. For example, the cases when we use the partition coefficient for the fuzzy measure, Algebraic product or Hamacher product when $p = 0.7$ for the *T*-norm, and function *g* defined in (7) are the best accuracy as 0.9937.

Classification accuracy of data having the value of reliability score is 0.99 or more at 1st candidate code

Table 2

Condition of reliability score (R.S.)						Number of total instances	Number of assigned instances at 1st candidate with R.S. ≧0.99	Number of matched instances at 1st candidate with R.S. ≧0.99	Classifiacation accuracy at 1st candidate with R.S. ≧0.99
Fuzzy measurement	Partition Entropy	T-norm	Algebraic prod.	$g(n_i)$	NA (Original)	40,000	30,089	28,532	0.9483
					Sigmoid func. (a)		20,544	20,413	0.9936
					Sigmoid func. (b)		24,971	24,598	0.9851
			Hamacher prod.		Sigmoid func. (a)		20,544	20,413	0.9936
					Sigmoid func. (b)		24,971	24,598	0.9851
	Partition Coefficient		Algebraic prod.		Sigmoid func. (a)		20,349	20,220	0.9937
					Sigmoid func. (b)		24,801	24,428	0.9850
			Hamacher prod.		Sigmoid func. (a)		20,349	20,220	0.9937
					Sigmoid func. (b)		24,801	24,428	0.9850

We also investigated the robustness of our classifier using the same dataset as the one used in the first experiment. Note that, we investigated the classifier that utilizing the partition coefficient as a fuzzy measurement to generalized reliability scores, excluding negative restriction for the simulation of the test of robustness for partition entropy as a fuzzy measurement. We performed code assignments under several conditions (See Table 3) with several reliability scores.

Conditions of code assignment

Table 3

No.	noise Δ	No.	noise Δ
1	NA	5	$\delta_{jk} \sim \mathrm{N}(0, 0.07^2)$
2	$\delta_{jk} \sim \mathrm{N}(0, 0.01^2)$	6	$\delta_{jk} \sim \mathrm{N}(0, 0.09^2)$
3	$\delta_{jk} \sim \mathrm{N}(0, 0.03^2)$	7	$\delta_{jk} \sim \mathrm{N}(0, 0.10^2)$
4	$\delta_{jk} \sim \mathrm{N}(0, 0.05^2)$		

Table 4 shows the classification accuracy of the classifier employed the generalized reliability score based Algebraic product and function g shown in (7) under each condition of σ. The number of instances that match with each candidate code gradually decreases as the value of σ becomes larger. However, it decreases very gradually, and it can be seen that generally, the solutions of the classifier with this generalized reliability score have robustness. Furthermore, table 5 shows the difference of the total classification accuracy compared to the normal classification under each condition of σ and each reliability score. From table 5, it can be seen that our classifier with the generalized reliability score has robustness. In particular, the results when we use the generalized reliability scores based on the Hamacher product have better performance compared with other generalized reliability scores using the Algebraic product, which are our original proposed reliability scores.

Classification accuracy under each condition

Table 4

	Number of total instcances	Number of matched instances						
		Noise $\Delta = (\delta_{jk}), \quad j = 1, \ldots, J, \quad k = 1, \ldots, K, \quad \delta_{jk} \sim N(0, \sigma^2)$						
		NA	$\sigma = 0.01$	$\sigma = 0.03$	$\sigma = 0.05$	$\sigma = 0.07$	$\sigma = 0.09$	$\sigma = 0.10$
1st candidate	40,000	35,118	35,108	35,123	34,972	34,842	34,644	34,135
2nd candidate		1,617	1,610	1,577	1,643	1,607	1,491	1,503
3rd candidate		540	542	556	495	387	400	461
4th candicate		292	276	248	210	192	204	247
5th candidate		178	173	156	134	144	139	176
Total		37,745	37,709	37,660	37,454	37,172	36,878	36,522
$d_{n\,total}$: difference of total accuracy		1.000	0.999	0.998	0.992	0.985	0.977	0.968

Difference of total accuracy compared to the normal classification under each condition

Table 5

		$d_{n\ total}$: difference of total accuracy						
		Noise $\Delta = (\delta_{jk}),\quad j = 1, \ldots, J,\quad k = 1, \ldots, K,\ \delta_{jk} \sim N(0, \sigma^2)$						
		NA	$\sigma = 0.01$	$\sigma = 0.03$	$\sigma = 0.05$	$\sigma = 0.07$	$\sigma = 0.09$	$\sigma = 0.10$
Choice of T-norm and Sigmoid func.	Algebraic Prod. / Sigmoid func(a)	1.000	0.999	0.998	0.992	0.985	0.977	0.968
	Algebraic Prod. / Sigmoid func(b)	1.000	0.999	0.998	0.992	0.985	0.977	0.968
	Hamacher Prod. / Sigmoid func(a)	1.000	0.999	0.998	0.993	0.985	0.979	0.974
	Hamacher Prod. / Sigmoid func(b)	1.000	0.999	0.998	0.993	0.986	0.980	0.974

6. CONCLUSION

This paper proposes a generalized reliability score based on the idea of T-norms in statistical metric space and the consideration of the frequency of each object (or feature) over codes in the overlapping classifier. In addition, we investigate the robustness of the overlapping classifier based on the generalized reliability score. As we often cannot control the quality of observed data that we handle, the generalization of the reliability score and the investigation of the robustness of the algorithm of the overlapping classifier are required for stable data processing. The numerical examples show that our overlapping multiclass classifier employing the generalized reliability score gives better results compared with the results of our previous studies. In addition, the investigation of the robustness of the classifier shows the stable code assignment of the classifier.

In future work, as the R package of our overlapping classifier is under development, we plan to release our classifier as an R package to contribute the coding tasks.

References

1. **Bezdek, J.C.**, 1981, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.
2. **Bezdek, J.C., Keller J., Krisnapuram, R., Pal, N.R.**, 1999, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer Academic Publishers, New York.
3. **Dowle, M., Srinivasan, A.**, 2019, *data.table: Extension of `data.frame`*, R package version 1.12.0., *https://CRAN.R-project.org/package=data.table*.
4. **Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., Steiner, S**., 2017, *"Three methods for occupation coding based on statistical learning"*, Journal of Official Statistics, Vol. 33, No. 1, pp. 101-122. DOI: https://doi.org/10.1515/jos-2017-0006.
5. **Hacking, W., Willenborg, L.**, 2012, *"Coding; interpreting short descriptions using a classification"*, Statistics Methods, Statistics Netherlands, The Hague, Netherlands, Available at: *https://www.cbs.nl/en-gb/our-services/methods/statistical-methods/throughput/throughput/coding* (accessed August 2018).

6. **Kudo, T., Yamamoto, K., Matsumoto, Y.**, 2004, *"Applying conditional random fields to Japanese morphological analysis"*, in the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25-26, Jul. 2004, pp. 230-237.
7. **Menger, K.**, 1942, *"Statistical metrics"*, Proceedings of the National Academy of Sciences of the United States of America, Vol. 28, pp. 535-537.
8. **Mizumoto, M**., 1989, *"Pictorical representation of fuzzy connectives, Part I: Cases of T-norms, t-Conorms and Averaging Operators"*, Fuzzy Sets and Systems, Vol. 31, pp. 217-242.
9. **Mullen, L. A., Benoit, K., Keyes, O., Selivanov, D., Armold, J.**, 2018, *"Fast, Consistent Tokenization of Natural Language Text"*, Journal of Open Source Software Vol. 3, No. 23: 655, https://doi.org/10.21105/joss.00655.
10. **Schweizer, S., Sklar, A.**, 2005, *Probabilistic metric spaces*, Dover Publications, New York.
11. **Shimono, T., Wada, K., Toko, Y.**, 2018, *"A supervised multiclass classifier using machine learning algorithm for autocoding"*, Research Memoir of Official Statistics, Vol. 75, pp. 41-60 (in Japanese).
12. **Toko, Y., Wada, K., Kawano, M.**, 2017, *"A supervised multiclass classifier for an autocoding system"*, Romanian Statistical Review, Vol. 4, pp. 29-39.
13. **Toko, Y., Wada, K., Iijima, S., Sato-Ilic, M.**, 2018a, *"Supervised multiclass classifier for autocoding based on partition coefficient"*, Czarnowski, I., Howlett, R.J., Jain, L. C., and Vlacic, L. (Eds.), Intelligent Decision Technologies 2018, Smart Innovation, Systems and Technologies, Springer, Switzerland, Vol. 97, pp. 54-64.
14. **Toko, Y., Iijima, S., Sato-Ilic, M.**, 2018b, *"Overlapping classification for autocoding system"*, Romanian Statistical Review, Vol. 4, pp. 58-73.
15. **Wickham, H., Francois, R., Henry, L., Muller, K.**, 2018, *dplyr: A Grammar of Data Manipulation*. R package version 0.7.8. https://CRAN.R-project.org/package=dplyr.