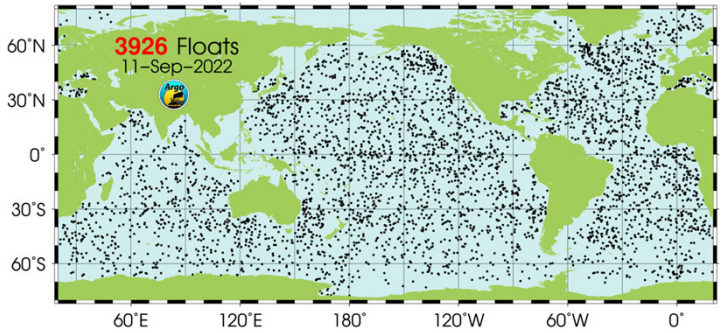
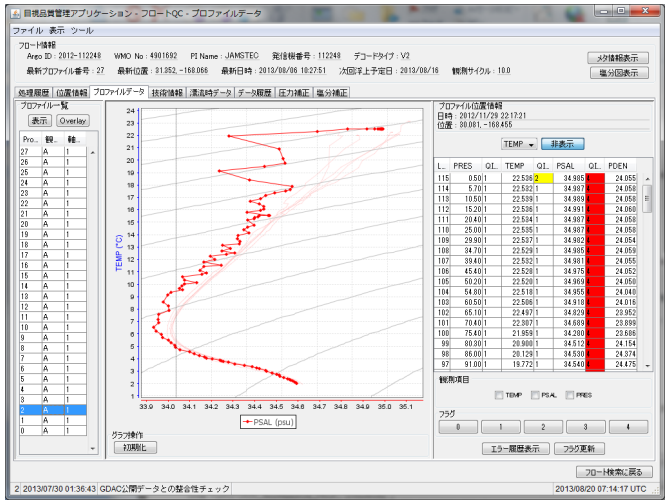
Argo real-time QC procedure using signature-based neural network

Shigeki Hosoda, Nozomi Sugiura, Shinya Kouketsu, Kanako Sato, Tadashi Hemmi
(JAMSTEC, RIGC)

Introduction

- More than 20 years since Argo programme was started, the huge number of profile data has been accumulated over 2.6M profiles
- Profile data are mostly correct, but error data with out of Argo's accuracy are occasionally recorded, being checked through r&dQC procedures
- Fully automatic QC procedures are difficult due to difficulties of dividing into natural variability, still requires human resources.
- Here, a method of error detection is provided using profile shape (=Signature) with neural network for judgement error data of Argo profiles





Signature method (Sugiura and Hosoda, 2020)

A single T, S, and P profile can be represented as iterated integrals ($\mathbf{X_k}$):

$$\mathbf{X}_k^{(i_1,i_2,\cdots,i_k)} = \int_{\tau_k=0}^{t} \cdots \int_{\tau_2=0}^{\tau_3} \int_{\tau_1=0}^{\tau_2} dX_{\tau_1}^{i_1} dX_{\tau_2}^{i_2} \cdots dX_{\tau_k}^{i_k}.$$

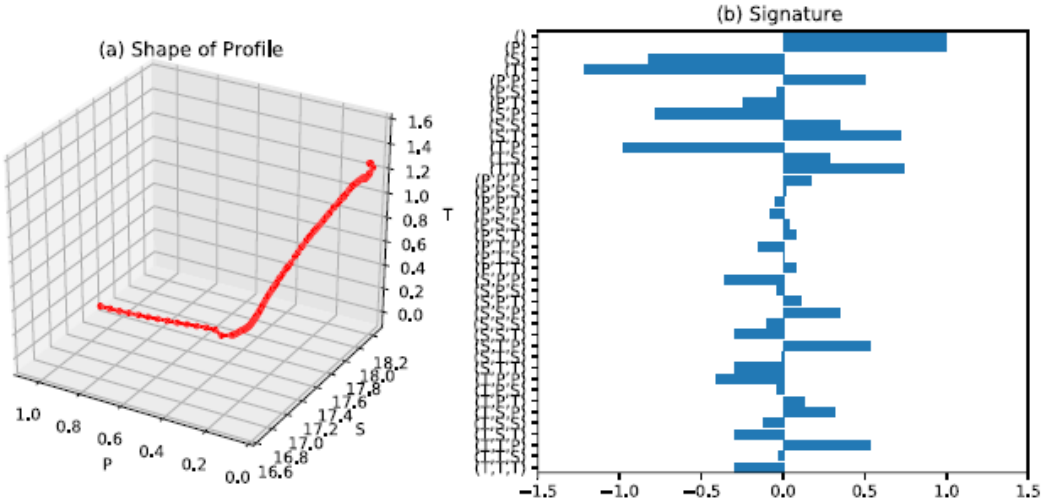
Here, each superscript i_k =1,2,3 represents an element like T, S, or P. Then, the signature up to degree n, is defined as

$$S_n(X) = (\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n).$$

In the simplest case, a function value for a profile with error is represented as a linear combination of the iterated integrals $\{\mathbf{X}_k^{(i_1,i_2,\cdots,i_k)}, k = 0, \cdots, n\}$:

$$y = \sum_{k=0}^{n} \sum_{i_1,i_2,\cdots,i_k} w^{(i_1,i_2,\cdots,i_k)} \mathbf{X}_k^{(i_1,i_2,\cdots,i_k)} + \varepsilon,$$

Here, *y* is the discrimination value,1 (incl.error) or 0 (correct data), $\{w^{(i_1,i_2,\cdots,i_k)}, k = 0, \cdots, n\}$ is a set of weights that minimizes a cost function with respect to the pairs of training data (*X(m), y(m)*), *m=1,⋯,M*. ε：error.



An example of a profile shape (a) and its signature S_n(X) (b). For example, (T,P) denotes the iterated integral $\mathbf{X}_2^{(T,P)} = \int_0^t \int_0^{t_2} dT_{t_1} dP_{t_2}$, P, T, and S in (a) are non-dimensionalized by dividing by 2000 dbar, 20degC, and 2psu, respectively (ref. Sugiura and Hosoda, 2020). Signature concept is introduced by “Rough Path Theory” (Lyons, 1998, Lyons et al., 2007, Lyons, 2015). The number of training data used here is 40% of profiles.

Signature-based neural network

The precision (rate of correct prediction for bad profiles) in the signature method was not high (see right Fig.).

How can we get a higher precision in the signature method?

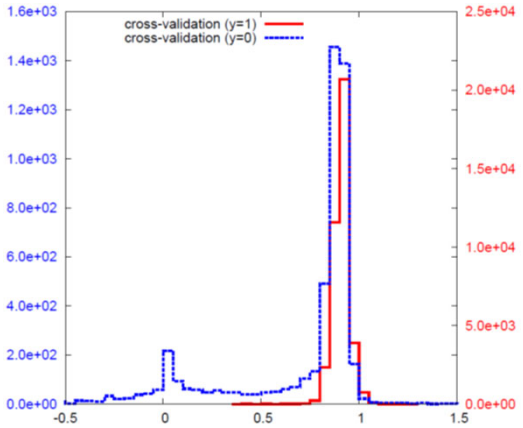


Fig. : Precision for good profile (red: y=1) is mostly judged around 1 (x-axis). However, precision for bad profile (blue: y=0) has a large peak around x=1 while small around y=0.

w^l in the signature method can be represented as linear combination. However, there are multiple factors on errors, like, domain segmentation, variable kinds of incorrect data.

Here we try to represent the weight function with a CNN (ResNet like) neural network for non-linearity, and try to improve with a metric learning methods (ArcFace; Deng et al., 2019) with Profile QC flags (AAA-FFF, 216 dimensions; AAA= good). *F(S(X))* is used instead of *w(S(X))* as follows.

$$F(S(X)) = FC1 \circ RN_{3,256\Rightarrow1,256} \circ RN_{7,256\Rightarrow3,256} \circ RN_{15,256\Rightarrow7,256} \circ Conv1(S(X))$$

$Conv1$:
CNN(*filter size = 4097x1, input:59987x1 → output:15x256, activation → ReLu*)

$RN_{ni,mi\Rightarrow no,mo}$ (X):
X_n = *MaxPool(ni → no)*(X)
CNN(3, *no x mi → no x mo, ReLu*) → CNN(3, *no x mo → no x mo, ReLu*) + X_n

$FC1$: *F(216x256) ・X(256x1) + B, with softmax function* (modified following ArcFace)

Profile QC flags (for T,S,P):
AAA: good for T, S, P
FFF: bad for T, S, P
* Chars A-F depending on the num of levels with flag=4.

Softmax: $f(x_i)=\exp(x_i)/\Sigma\exp(x_i)$,
ReLU: f(x)=0 (x<0), f(x)=x (x>0).

Results

Here we use sample data for training and testing in the case of JMA

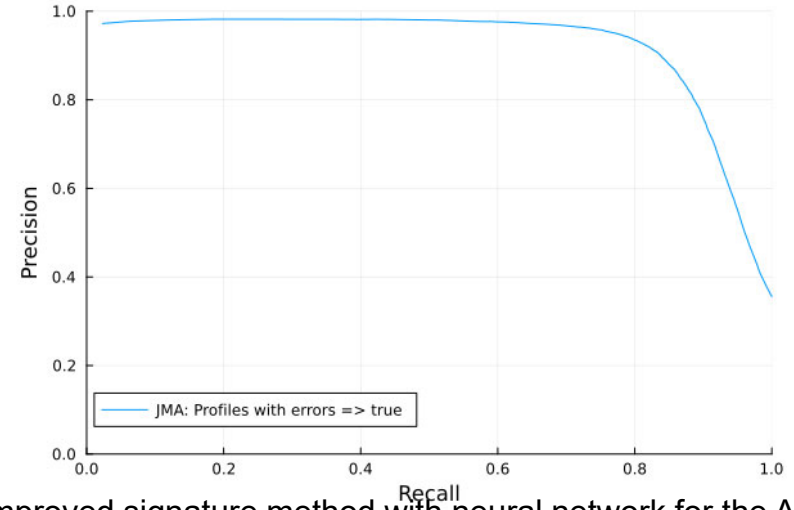
- Pick up profiles with N_LEVELS >2 and P_max-P_min > 1000 dbar

Judgement definition

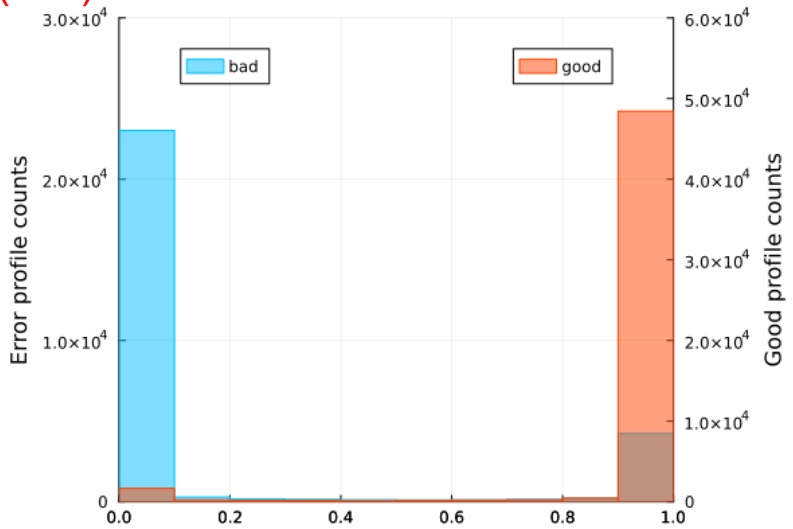
- Define “bad profile data” if at least one level with flag=4 is included in PRES_ADJUSTED_QC or TEMP_ADJUSTED_QC or PSAL_ADJUSTED_QC

The results in the signature method are improved: Precision is 0.4 → 0.9 @ Recall is 0.8.

There are peaks clearly separated for y=1 (good) and y=0 (bad)



Improved signature method with neural network for the Argo profile (French DAC; JMA). Here we show a relationship between Precision and Recall (rate of correct prediction that is judged as the same as original QC).

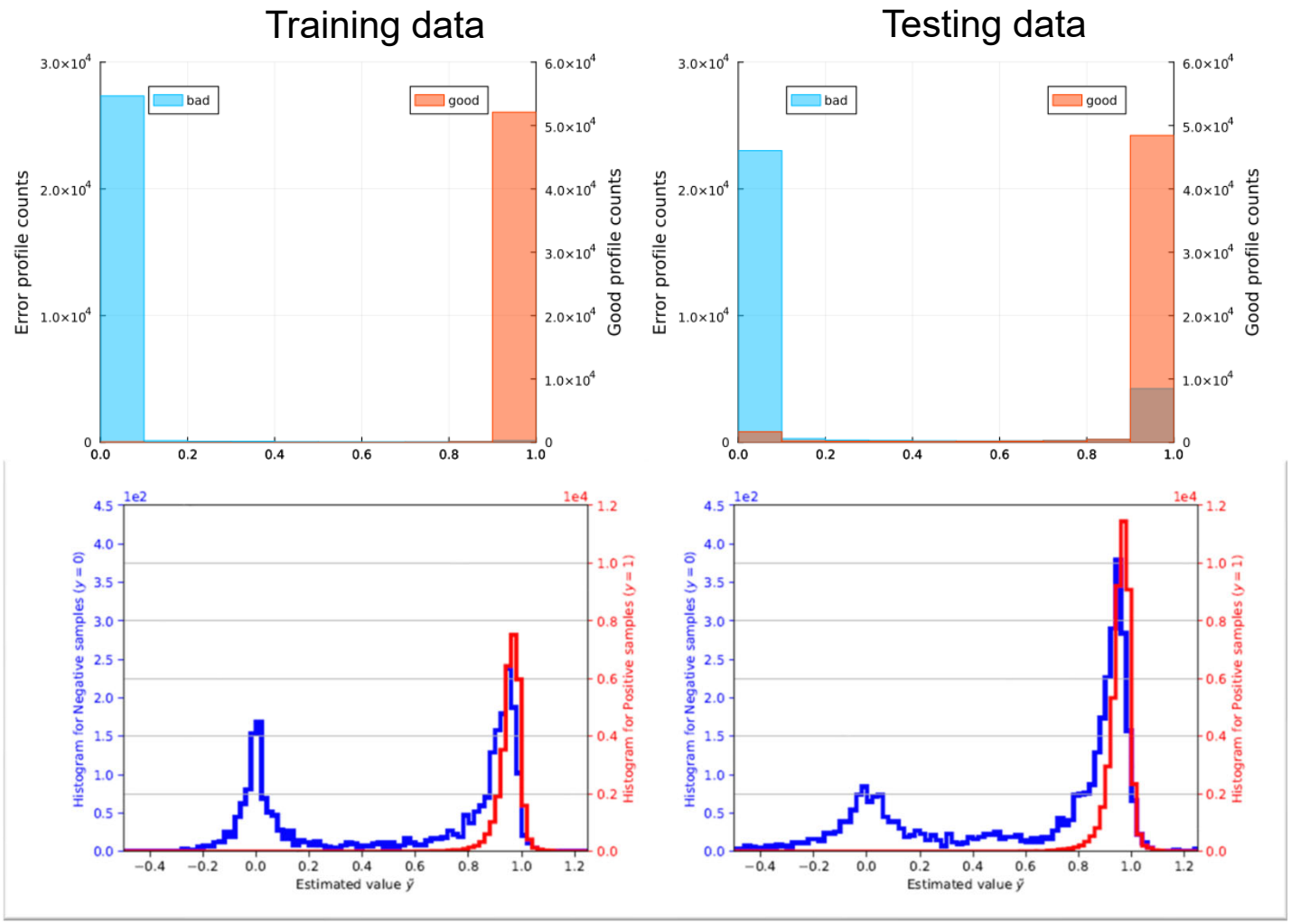


Improved results (same as left case) : frequency distribution of discriminant values for the good (light blue) and bad (purple) profiles.

Comparison with previous method

New

Old



Histograms of discriminant standard values. (Top) Improved signature method, (Bottom) previous version. (Left) Training data, (Right) Testing data. The red/orange lines indicate a profile that the correct answer is “good" and the blue/light blue lines indicate a profile that the correct answer is “bad".

The new version has two peaks around y=1 and 0 for good and bad data.
→ Cleary improved by introducing the CNN (ResNet like) neural network for non-linearity.

Summary and Conclusion

- Signature method is improved to detect error data by neural networks that can represent nonlinearity into the weight w^l.
- Improved result is, Precision is 0.4 → 0.9 when Recall is 0.8, in the case of Japanese DAC (JMA).
- By using the neural network technique to *w*, nonlinear information of errors (sea area, error type etc.) can be represented in the profile information, which is efficiently expressed with the signature *(S(X)).*

More testing is needed to apply real QC procedure and to improve recall, e.g., need to input detailed information in the signature method based on lots of experiences in dQC process.

This work was supported by JST, AIP Trilateral AI Research, Grant Number JPMJCR20G5, Japan.