Decision Tree-based Feature Function Design in Conditional Random Field Applied to Error Detection of Ocean Observation Data

Yosuke Kamikawaji and Haruki Matsuyama
Department of Information Science and Biomedical Engineering, Graduate School of Science and Engineering, Kagoshima University
Kagoshima, Japan
Email: {sc112015, sc109066}@ibe.kagoshima-u.ac.jp

Ken-ichi Fukui
The Institute of Scientific and Industrial Research, Osaka University
Suita, Osaka, Japan
Email: fukui@ai.sanken.osaka-u.ac.jp

Shigeki Hosoda
Japan Agency for Marine-Earth Science and Technology
Yokosuka, Kanagawa, Japan
Email: hosodas@jamstec.go.jp

Satoshi Ono
Department of Information Science and Biomedical Engineering, Graduate School of Science and Engineering, Kagoshima University
Kagoshima, Japan
Email: ono@ibe.kagoshima-u.ac.jp,

***Abstract*—Globally-covered ocean monitoring system Argo with more than 3,700 floats has been working, and its accumulated big ocean observation data helps many studies such as investigation into climate change mechanism. Since the observed data sometimes involves errors, human experts must visually confirm and revise quality control (QC) labels. However, such manual QC by human experts cannot be performed in some contries. In addition, it is difficult to regularize the quality of the ocean observation data of all over the world because the manual QC depends on human experts' heuristics. Therefore, this paper proposes a method for error detection in Argo observation data using Conditional Random Field (CRF) to realize an automatic QC with high accuracy equal to human experts. This paper also proposes a feature function design method using decision tree learning. Using decision tree allows coping with various types of observation errors without manual work, whereas previous work had to focus on certain error types due to manual labor for feature function design. Experimental results have shown that the proposed method could detect all types of salinity errors with autoamtically designed features while maintaining the higher accuracy of QC label assignments than the actually operated system in Argo project and a previous method using CRF with SVM.**

1. Introduction

The ocean is regarded as a driving source of the global warming and climate changes due to its heat capacity, 1,000 times as much as of air, though their mechanisms have not been clarified sufficiently yet. Although the physical state observation of the ocean on a regular basis anywhere in the world is necessary to understand oceanic variability, it is difficult by research vessels to conduct such ocean monitoring. Therefore, the international Argo program has started since 2000 [1], [2], [3], aiming to realize an autonomous, long-term, real-time and globally-covered ocean monitoring system. In this program, over 30 countries of all over the world have constructed an array of over 3,700 "Argo floats", which reports observation data in real time via satellites. The observed data is opened via the Internet within 24 hours after received from the float, and is utilized for a weather forecast and various industries such as sea transportation and fishing in addition to many studies about climate change mechanism [4], [5], [6]. Thanks to produce lots of the Argo float data, accuracy of forecasting long-term/large scale climate change such as El Nino and La Nina and Indian Ocean Dipole Mode has been largely improved [7].

Argo float sometimes fails in observation due to the hardware or software, as well as external factors such as sensor contamination and data reception failure. It is difficult to evaluate the data accuracy obtained from Argo float after being thrown into the ocean. Therefore, "quality control" (QC) is performed that finds and corrects observation error according to enormous accumulation of observation data. In this QC, a QC flag is assigned to each observed value. However, for automated QC, the data accuracy is not as high as it can withstand research needs. This is due to the naturally occurring variation in water temperature and salinity, where the S/N value between the variation amount and error is not so high, making it difficult to automatically distinguish between the error and the signal. As a result, overlooked error (failure to detect error) and misdetection (extraction failure) occur with the current automated QC. Furthermore,

it becomes troublesome for technicians to manually performing visual check. In addition, among technicians, there exist differences in correction standards, not enough human resource, or countries with unskilled workers, thus unable to have uniform global quality. This has been a problem over long period of time among international Argo project, and it is affecting the accuracy and reliability of the global ocean environment monitoring.

Aiming to improve the data accuracy and quality uniformity in the entire global region, this study proposes a method to automatically classify the QC flag of the ocean data. This work modeled a problem of error detection and QC flag assignment of the ocean data as a sequential labeling problem, and performs classification using Conditional Random Field (CRF) [10].

Applying CRF to QC flag assignment requires feature selection and threshold adjustment because the observation data involves continuous values and the error pattern depends on observation depth. In [9], appropriate explanatory variable combinations were designed by interview to technicians to handle the latter difficulty, and an automatic threshold adjustment mechanism is introduced to deal with the former. In detail, the previous method utilized Support Vector Machine (SVM) as a discriminant function in feature conditions of CRF, relaxing the difficulty of threshold adjustment for every observation ocean area. However, this method still requires human experts to clarify a combinations of explanatory variables to design features for each observation error type and ocean area. In addition, it is difficult even for human experts to categorize the observation errors in Argo data. Then, the previous work had to focus only on a few typical observation types such as density inversion.

Therefore, this paper proposes a method for designing feature functions rather than the threshold adjustment. The proposed method utilizes Decision Tree (DT) learning to choose a proper explanatory variable combination and their thresholds, which must be designed for each observation error type.

The main contributions of this paper, which also represents the progress of this paper from the previous work [8], [9], are as follows:

- A sequential labeling method using CRF for Argo float data error detection, which requires consideration of continuity in both viewpoints of features and output data quality labels.
- A decision tree learning-based feature function design method, which simultaneously identifies an appropriate combination of explanatory variables and proper thresholds for each of them. This permits the application of the proposed method to all types of observation errors without manually designing feature functions.
- Comparative study of the proposed sequential labeling method using DT learning with the previous method using SVM and the actually operated system in Argo project (real-time QC), which reveals that the proposed method showed better accuracy for detecting errors spreading over multiple layers than the previous method, in particular.

2. Global ocean monitoring system Argo

2.1. Ocean observation by Argo floats

Fig. 2 shows the measurement cycle with the Argo floats. After being released into the ocean, the Argo floats drift to water depth 1,000 [m], where they are less affected by the ocean current. When the observation time comes, it descends to water depth 2,000 [m] then ascends while measuring water temperature and salinity. The measurement data created during one floating is called a profile, and each profile records 100 vertical layers of temperature and salinity values. Argo floats automatically carry out this measurement cycle every 10 days. Fig. 4 shows an example of the profile. Vertical axis represents pressure [dbar], which is almost the same as water depth [m]. Blue graph represents the water temperature [$^\circ$C], red graph the salinity [PSS-78], and the green graph the density $[\mathrm{kg/m^3}]$ obtained from the temperature and salinity. Sensor accuracies of pressure, temperature and salinity are necessary within ± 2.5 [dbar], $\pm 0.005^\circ$C and ± 0.01 [psu], respectively, which have been decided by Argo Data management Team [3].

When the observed data is sent to ground station via a satellite, it goes through quality control and released on the Internet. During the quality control of the Argo, the reliability of the measurement value is set for the profile measured by the Argo floats following above accuracy criteria. As quality control flags, 4 levels of reliability is used: 1 (correct), 2 (probably correct), 3 (probably incorrect), and 4 (incorrect).

There are 2 types of quality control: *Real-time quality control (RQC)* and *Delayed-mode QC (DQC)*. RQC is a simple quality control that aims to release the measured data within 24 hours of profile observation. Since it puts priority on the real-time release of data, visual checking by technicians is often not done. On the other hand, the data qualified through DQC is used for research purposes to be further analyzed, so the quality control is done with precision. They are corrected through visual check by technicians.

The observation value is greatly affected by the naturally occurring changes in temperature and salinity, thus the influence to the observation value by errors becomes small. This is due to seasonal and climate changes affecting the water temperature and salinity. Such natural trends make it difficult to describe the data QC method. Although the existing automated RQC method has detailed measurement method regulations, it is technically difficult to correspond to all of the errors, resulting in the overlook of error or false detection. Ultimately, it is checked visually and corrected manually by the technicians but this is becoming a big burden on technicians.

In addition, when human decision is involved, it leads to a collapse of uniformity in data quality. It is quite difficult

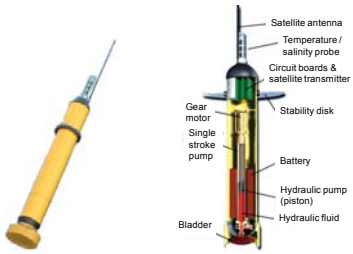


Figure 1. Argo float.

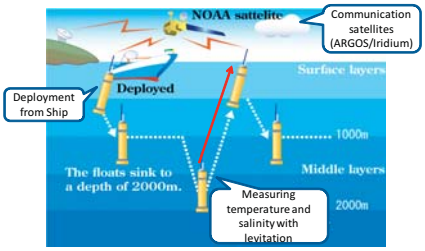
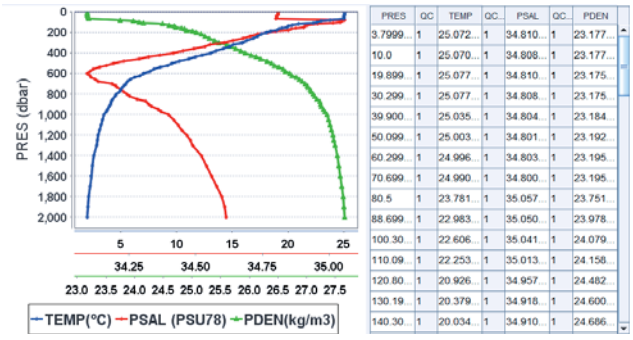


Figure 2. Observation cycle.

Figure 3. Argo float distribution map[3].



PRES	QC	TEMP	QC	PSAL	QC	PDEN
3.7999...	1	25.072...	1	34.810...	1	23.177...
10.0	1	25.070...	1	34.808...	1	23.177...
19.899...	1	25.077...	1	34.810...	1	23.175...
30.299...	1	25.077...	1	34.808...	1	23.175...
39.900...	1	25.035...	1	34.804...	1	23.184...
50.099...	1	25.003...	1	34.801...	1	23.192...
60.299...	1	24.996...	1	34.803...	1	23.195...
70.699...	1	24.990...	1	34.800...	1	23.195...
80.5	1	23.781...	1	35.057...	1	23.751...
88.699...	1	22.983...	1	35.050...	1	23.978...
100.30...	1	22.606...	1	35.041...	1	24.079...
110.09...	1	22.253...	1	35.013...	1	24.158...
120.80...	1	20.926...	1	34.957...	1	24.482...
130.19...	1	20.379...	1	34.918...	1	24.600...
140.30...	1	20.034...	1	34.910...	1	24.686...

Figure 4. Example profile.

to guarantee the data uniformity required for global Argo observation network. This is due to the fact that there are differences in correction standards among technicians, and the presence of countries where they lack correction technicians. Furthermore, if a modification of QC method occurs, one must reconfirm the previously QC applied profiles but the lack of human resource makes this difficult. This has been a big problem for the international Argo project, which affects the accuracy and reliability of global ocean environment monitoring system.

2.2. Error types in Argo observation data

Table 1 shows all type errors of Argo observation data related to salinity. Note that it is not based on oceanographical knowledge but expediential classification in this study[1]. The profiles shown in the table were observed in North Pacific Ocean that were maintained through QC in Japan.

Previous method [8], [9] targeted only error types marked with '*', and successfully detected and assigned QC labels to those errors. However, observation error types must be clarified and feature functions for error types must be designed manually in advance. It is quite difficult even for experts to define observation error types and categorize them. Furthermore, feature function design and threshold adjustment are required for each ocean area and error type.

2.2.1. Density inversion. *Density inversion* is an observation error that occurs from substance contamination or adhesion of organisms. Density is the value calculated from water temperature and salt content and it increases monotonically along with the depth in most sea, thus we can judge that an error has occurred if the density is vertically reversed.

Therefore, the observation error can be detected by detecting vertical reversal relations up to a certain threshold. Density inversion often occurs when observation error is found in either water temperature or salinity, and is determined by the relationship of the value of the upper and lower layers. In order to ultimately determine the error, visual verification by a specialist is needed. Since it includes naturally occurring fluctuations, distinction between observational error and natural fluctuation is needed and is difficult.

2.2.2. Equivalence error. When Argo floats make observations, the previously observed value may be copied if there is a voltage shortage. In other words, repeated observation value is stored at continuous layers. In this paper, this type of observation error is called *equivalence error*. However, it is permitted if there is a continuous small difference in water pressure difference between the observation layers, where flag 4 is not assigned.

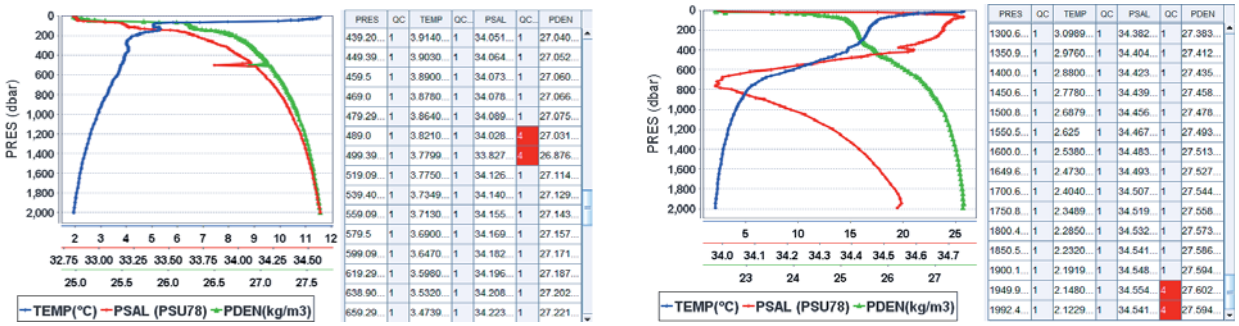
2.2.3. Errors over the whole profile. In Argo data, QC labels are assigned to observation layers and sometimes observation errors over multiple layers occur. Errors over the whole profile involve more than 70% layers marked with error labels, and are categorized into some detailed error types such as offset, zigzag, and so on. *Offset error* occurs when compared with past profile or with neighboring profile, and the observables seem to have shifted horizontally as shown on Fig. 7, where thin lines indicate other profile previously observed by the same float. Offset sometimes occurs in the whole observation layer, or only in the deep layer, thus it is considered detection exception with RQC. *Zigzag error* involves perturbed observation values as shown in Fig. 6. *Sensor failure* is a kind of errors that are continuously occurred on the same float possibly due to a hardware problem. In contrast, *external factor* is an error whose preceding and following profiles observed by the same float do not involve observation errors possibly due to an external factor.

2.2.4. Other errors. Other types of errors are classified as follows: *Pressure errors* are salinity errors caused by errors on pressure value. *Errors in shallow depth* are errors in the

1. We found none of such taxonomies on Argo observation error.

TABLE 1. EXPEDIENTIAL CLASSIFICATION OF ARGO OBSERVATION ERRORS.

Error type	Number of profiles	Detailed type	Number of profiles
Density inversion	1,253	Negative*	1,039
		Positive*	203
		Both	11
Equivalence error*	220	———	
Mix of density inversion and equivalence error*	89	———	
Outlier*	352	———	
Error over the whole profile	1,494	Zigzag	63
		Offset*	660
		Sensor failure	223
		External factor	3
		Others	545
Pressure error	1,282	———	
Error in shallow depth	2,156	———	
Too few layers	1,655	———	
Label 9 (missing value)	328	———	
Unknown reason	134	———	
Total	8,963	———	



(a) Example negative density inversion error

(b) Example density inversion error in deeper depth

Figure 5. Density inversion error

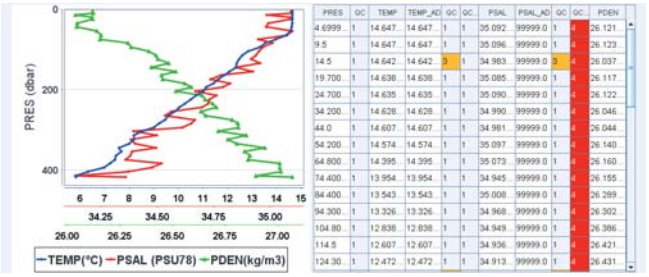


Figure 6. Example error over the whole profile (zigzag).

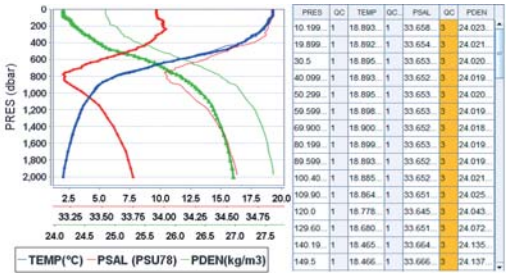


Figure 7. Example offset error.

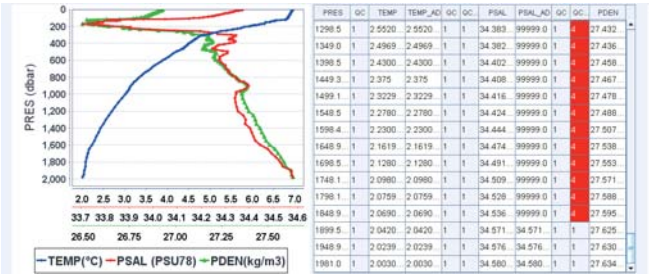


Figure 8. Example error over the whole profile (external factor).

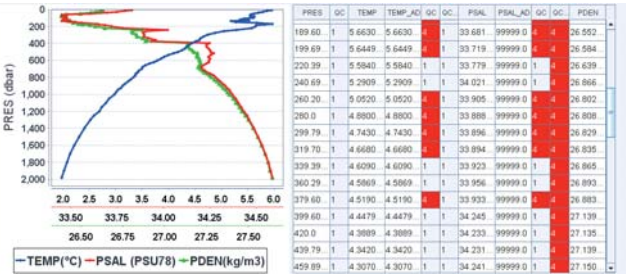


Figure 9. Example unknown error.

shallowest or nearby layers. *Too few layers errors* are ones in a profile involving less than 10 observation layers. *Errors with unknown reason* are ones that cannot be categorized into the above error types as shown in Fig. 9.

2.3. Difficulty to design feature functions for Argo QC label assignment problem

In general, CRF is applied in tasks with symbolic attributes such as natural language processing, where explanatory variables are part of speech, mora length, and so on. However, ocean observation data is numeric, continuous attributes, and essentially feature functions involve inequation conditions. Most of inequation conditions involve thresholds for explanatory variables that must be adjusted depending on error types and ocean areas.

In addition, a feature condition must be a combination of explanatory variables rather than single variable because there are dependencies between error pattern and sea depth. For instance, density inversion with large inversion width is sometimes not regarded as error in shallower depth, whereas the ones with slight width must be error in deeper depth. This is because, at the shallower depth, salinity and temperature greatly change compared to deeper depth such as 800 meter from the sea surface or more because of seasonal influence and other factors.

Here, a simple preliminary experiment was conducted to clarify the necessity of considering combinations of explanatory variables in a feature function. Profiles involving negative density inversion were used in this experiment, and two methods were compared; the first one is the method proposed in [8] that detects negative density inversion with a feature function comprising four conditions as follows

$$\begin{aligned} PDEN_t - PDEN_{MAX} < -0.01 \quad \wedge \quad PRES_t <= 1,400 \\ PDEN_t - PDEN_{MAX} < 0 \quad \wedge \quad PRES_t > 1,400 \end{aligned} \tag{1}$$

where $PRES_t$ and $PDEN_t$ indicate sea depth of observed layer t and density value at t, and $PDEN_{MAX}$ denotes the maximum value of the density at the depth range from 1 to $t-1$. The second method is almost the same as the first method except the feature functions for negative density inversion; the following four functions comprising each condition are used instead of eq. (1).

$$PDEN_t - PDEN_{MAX} < -0.01 \tag{2}$$
$$PDEN_t - PDEN_{MAX} < 0 \tag{3}$$
$$PRES_t <= 1,400 \tag{4}$$
$$PRES_t > 1,400 \tag{5}$$

That is, the difference between two methods was whether the dependency between four conditions was considered or not. Other experimental conditions were configured the same as the previous work [8].

Table 2 shows the number of layers of test profiles. As shown in the table, the second method overlooked many errors because it is difficult to consider the dependency between the error patterns and observation depth.

TABLE 2. PRELIMINARY EXPERIMENT

Methods	Precision	Recall
Method 1 (features considering explanatory multiple variables)	0.94	0.81
Method 2 (features involving one variable)	0.97	0.54

3. The proposed method

3.1. Overview

This paper proposes an error detection method for Argo float data based on the following key ideas:

Idea 1: Modeling as sequential labeling and applying Conditional Random Field (CRF) [10]. This paper models a QC flag classification task in Argo data as sequential labeling, and proposes a method that utilizes CRF. CRF simultaneously considers various features, which is its advantage compared to Hidden Markov Model (HMM) [11]. The proposed method refers multiple features considering explanatory variable combinations, and assigns QC labels with considering combinations among neighboring labels.

Idea 2: Designing feature functions using decision tree learning. In previous work [8], [9], feature functions are designed by hand according to interview to technicians, and the designed feature functions worked well for detecting certain types of errors such as density inversion. As shown in Sec. 2.3, feature functions must consist of combinations of equation and inequation referring multiple explanatory variables because error patterns depend on water depth. This makes it difficult to design feature functions even for experts. In addition, feature functions must be designed for each ocean area and error type.

Therefore, the proposed method designs the feature functions in a fully automatic manner except determining the default rule. Decision tree learning allows designing a set of feature functions that consists of conditions of explanatory variable with a threshold. In addition, sequential labeling with the designed feature functions enables detecting continuous errors across multiple observation layers, while it is difficult for pointwise labeling method to detect them.

3.2. Conditional Random Field (CRF)

CRF is a discriminant model for sequence labeling proposed by Lafferty et al. [10]. Since sequential observation data and dependency with precedent and/or successive labels are available in Argo data, various features effective for label estimation can be implemented in CRF. This study adopts Linear-chain model, one of the simplest CRF model.

Given an input data $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$ whose length is T, conditional probability of $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)$ is modeled as follows:

$$P(y_t, y_{t-1}|\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{x}}} \exp\left(\sum_{t=1}^{T}\sum_{k} \lambda_k f_k(\boldsymbol{x}, y_t, y_{t-1})\right) \tag{6}$$

where $f_k(\boldsymbol{x}, y_t, y_{t-1})$ denotes a feature function that associates input data, t-th and $(t-1)$-th output labels, λ_k is a weight parameter of feature function $f_k(\boldsymbol{x}, y_t, y_{t-1})$, and $Z_{\boldsymbol{x}}$ is a regularization term that ensures $\sum_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}) = 1$. The feature function f_k gives features of input data to a classifier, and depends on input $\boldsymbol{x}$, labels y_t and y_{t-1} as follows:

$$f_k(\boldsymbol{x}, y_t, y_{t-1}) = \begin{cases} \phi_k & \text{if condition} = \text{true} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where ϕ_k is a real number called feature value, which is a constant determined according to f_k. Feature functions are designed to associate attributes of the target dataset and a possible labels of y_t and y_{t-1}.

In training phase, CRF calculates λ_k by maximum likelihood estimation from training data $\boldsymbol{D} = \{(\boldsymbol{x^{(i)}}, \boldsymbol{y^{(i)}})\}_{i=1}^{N}$ in which input data $\boldsymbol{x}$ and output label $\boldsymbol{y}$ are paired. When classifying unknown input data $\boldsymbol{x}$, output labels $\boldsymbol{y_{out}}$ can be determined by maximizing the following objective function $L(\boldsymbol{D})$:

$$L(\boldsymbol{D}) = \sum_{\boldsymbol{D}} \log P(y_t, y_{t-1}|\boldsymbol{x}) - \sum_k \frac{\lambda_k^2}{2\sigma^2} \quad (8)$$

where the second term is a regularization term used to prevent overfitting. Weight parameters λ_k are obtained by solving the above optimization problem using a steepest descent method.

When classifying unknown input data $\boldsymbol{x}$, output labels $\boldsymbol{y_{out}}$ can be determined by solving the following maximization problem:

$$\boldsymbol{y_{out}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} P(\boldsymbol{y}|\boldsymbol{x}) \quad (9)$$

3.3. Feature function design using decision tree learning

The proposed method identifies the QC flag by machine learning from observed profile of Argo. With the QC flag, we assign 1 (correct), 2 (probably correct), 3 (probably incorrect), or 4 (incorrect). However, since labels 2 and 3 include ambiguity, we consider 2 as 1, and 3 as 4 to model the problem as binary classification.

In this paper, we consider salinity QC flags as identification target, and try to detect all types of observation errors except ones marked label 9 (missing value), whereas previous work [8], [9] detects four errors: density inversion (positive and negative), equivalence error, outlier, and offset. The error types that were not handled in previous work have ambiguity tendency, making it difficult to manually design feature functions. Therefore, technique to design feature functions is mandatory to cope with all types of observation errors.

Decision tree is a kind of classification and regression models, and classify a class belonging an input or predict a value of one numerical variable. A decision tree has a root node that is a starting point of reasoning, non-terminal nodes (including the root node) that have outgoing edges to split data entries according to a value of the attribute associated with the node, and terminal nodes representing reasoning results, *i.e.*, class labels (classification) or estimated values (regression). A route to a terminal node consisting of a set of non-terminal nodes represents a feature set of the class of the terminal node; *i.e.*, the same number of feature functions as the number of terminal nodes are designed.

Constructing a decision tree for classification is performed by recursively selecting an explanatory variable that divides learning data most finely. Gini impurity is one of widely used index and defined as follows:

$$I_g = 1 - \left\{P(c_1|v = a_1)^2 + P(c_2|v = a_2)^2\right\} \quad (10)$$

where $P(c_1|v = a_1)$ and $P(c_1|v = a_1)$ are occurrence probabilities of groups that have c_1 and c_2 of an objective variable when they have values a_1 and a_2 of explanatory variable v. Evaluation of split condition is based on the change on I_g between before and after the split. The explanatory variable with the lowest I_g value is selected as variable at the current non-terminal node.

4. Evaluation

4.1. Experimental setting

To verify the effectiveness of the proposed method, experiments were conducted using the data observed in North Pacific Ocean area and maintained through QC in Japan that comprises 8,223 profiles. The profiles involved all kind of observation errors related to salinity except errors with label 9 that represents missing value, whereas previous work [9] tested four types of errors. The experiments were performed in 10-fold cross validation manner, and, while testing, we added the same amount of normal profiles without any observation errors as error profiles for test. In other words, 7,401 items of learning data, 1,644 items of prediction data (including profiles with observation errors: 822, normal profiles: 822). Although 10 experiments were performed while replacing with prediction data, this paper shows one of 10 holds because of no significant difference among 10 results.

In this experiment, we test the labeling of QC flags for observation value of salinity. As evaluation criteria, this paper focuses on the numbers of observation layers involving overlooked error and misdetection, respectively; we consider for layers including observation errors; TP is the number of layers in which errors were able to be correctly detected, and FN as the number of layers in which errors could not be detected. For the normal layers, TN is the number of layers that correctly detected no errors (correctly regarded as normal layers), and FP as the number of layers in which errors were mistakenly labeled as flag 4.

4.2. Feature function design using decision tree learning

Feature functions in the proposed method were designed using decision tree learning with the learning profile as

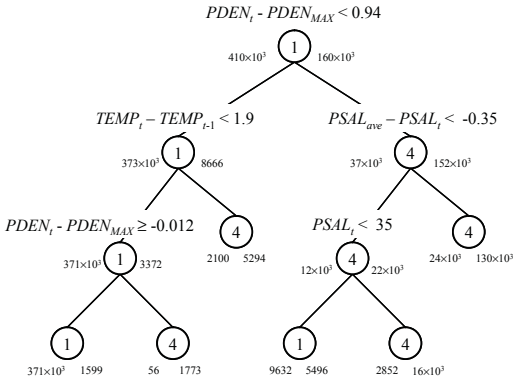


Figure 10. Decision tree obtained in the experiment.

described in Sec. 3.3 In this experiment, Classification And Regression Trees (CART) [12] is used and produced the tree shown in Fig. 10. In the figure, $PSAL_t$, $TEMP_t$, $PDEN_t$ denotes salinity, water temperature, and density at observation layer t. $PDEN_{max}$ denotes the maximum density until $(t-1)$-th layer, and $PSAL_{ave}$ indicate averaged salinity value over the whole profile.

Then, feature functions shown in Table 3 were obtained by combining all conditions of nodes from root to terminal nodes. f_1 is the only feature function that is manually designed to work as a default rule corresponding to valid observation values.

4.3. QC label assignment performance

The proposed method using the designed feature functions ($\mathrm{CRF_{DT}}$) as described in Sec. 4.2 was compared with RQC and previous method [9] using manually designed feature functions with SVM ($\mathrm{CRF_{man}}$) shown in Table 4. CRF used in this experiment was Linear-chain model. The learning of weight values was iterated until the sum of updated value became less than 10^{-10}. Point-wise labeling methods with the automatically designed feature functions ($\mathrm{PW_{DT}}$) and manually designed ones with SVM ($\mathrm{PW_{man}}$) were also compared. In point-wise labeling methods, an error label is assigned when at least one of feature conditions except f_1 was satisfied. The point-wise labeling method with the automatically designed feature functions corresponds to a method directly using the decision tree shown in Fig. 10.

Table 4 shows the manually designed feature functions [9]. The features were designed to detect four errors: density inversion (positive and negative), equivalence error, outlier, and offset. f_5 and f_6 had SVM classifier as discriminant functions whose input was pressure value and density inversion width. In addition, data in World Ocean Atlas 2005 was used to detect offset errors in f_7, f_8, and f_9; $\overline{PSAL_t^{(WOA)}}$ and $\sigma(PSAL_t^{(WOA)})$ denote average WOA observation data at depth t and its standard deviation.

Table 5 shows the comparison results on QC label assignment accuracy. $\mathrm{CRF_{DT}}$ and $\mathrm{PW_{DT}}$ found various types of observation errors, whereas $\mathrm{CRF_{man}}$ and $\mathrm{PW_{man}}$ overlooked many errors. In particular, the recall rate of $\mathrm{CRF_{DT}}$ outperformed that of RQC whereas the precision rate were almost competitive. Comparison between $\mathrm{CRF_{DT}}$ and $\mathrm{PW_{DT}}$ revealed that the necessity of sequential labeling for detecting errors across more than one layers.

4.4. Discussion

Although the proposed method outperformed other methods for the dataset involving all types of observation errors, the performance of the proposed method for certain types of errors focused in previous work [8], [9] was deteriorated. Table 6 shows the results only for the four types of errors marked in Table 1. The proposed method $\mathrm{CRF_{DT}}$ was worse than not only $\mathrm{CRF_{man}}$ but $\mathrm{PW_{DT}}$. This is because density inversion and equivalence errors occur at one or a few layers and $\mathrm{CRF_{DT}}$ prioritized other types of errors occurring across greater number of layers. Techniques coping with data imbalanceness such as adaptive tuning of feature values or under/over sampling would increase the accuracy of the proposed method.

5. Conclusions

This paper proposed a method to assign QC labels to the ocean observation data. The proposed method uses DT learning for feature function design and CRF to sequential labeling. The experimental results showed that the proposed method found all types of observation errors with automatically designed feature functions, whereas the previous work focused only on certain types of errors and requires manual feature selection according to interview to technicians. In particular, the proposed method significantly outperforms the existing system RQC from the viewpoint of recall rate.

As future work, we plan to apply the proposed method to data observed at other sea area, and to introduce some techniques to manage imbalanceness between error types.

Acknowledgments

The authors would like to thank Ocean Circulation group, Research and Development Center for Global Change, JAMSTEC for their help. This study was partially supported by the Kurata Memorial Hitachi Science and Technology Foundation, by Takahashi Industrial and Economic Research Foundation, and by JSPS KAKENHI Grant Number 16K12490.

References

[1] D. Roemmich, O. Boebel, Y. Desaubies, H. Freeland, K. Kim, B. King, P.-Y. Le Traon, R. Molinari, B. W. Owens, S. Riser et al., "Argo: The global array of profiling floats," in Observing the Oceans in the 21st Century, C. J. Koblinsky and N. R. Smith, Eds. GODAE Project Office, Bureau of Meteorology, 2001, pp. 248–258.

[2] A. D. M. Team, "Report of the argo data management meeting," in Proc. Argo Data Management Third Meeting, Marine Environmental Data, 2002.

TABLE 3. FEATURE FUNCTIONS OBTAINED USING DECISION TREE LEARNING.

No.	Condition	Feature value
f_1	$2 \leq PSAL_t \leq 41$	$1/T$
f_2	$PDEN_t - PDEN_{max} < 0.94 \wedge TEMP_t - TEMP_{t-1} < 1.9 \wedge PDEN_t - PDEN_{max} \geq -0.012$	1.0
f_3	$PDEN_t - PDEN_{max} < 0.94 \wedge TEMP_t - TEMP_{t-1} < 1.9 \wedge PDEN_t - PDEN_{max} < -0.012$	1.0
f_4	$PDEN_t - PDEN_{max} < 0.94 \wedge TEMP_t - TEMP_{t-1} \geq 1.9$	1.0
f_5	$PDEN_t - PDEN_{max} \geq 0.94 \wedge PSAL_{ave} - PSAL_t < -0.35 \wedge PSAL_t < 35$	1.0
f_6	$PDEN_t - PDEN_{max} \geq 0.94 \wedge PSAL_{ave} - PSAL_t < -0.35 \wedge PSAL_t \geq 35$	1.0
f_7	$PDEN_t - PDEN_{max} \geq 0.94 \wedge PSAL_{ave} - PSAL_t \geq -0.35$	1.0

TABLE 4. FEATURE FUNCTIONS USED IN THE PREVIOUS METHOD [9].

No.	Target error type	Condition	Feature vaule ϕ_k
f_1	Normal	$2 \leq PSAL_t \leq 41$	$1/T$
f_2	Other	$PSAL_t < 2, 41 < PSAL_t$	1.0
f_3	Equivalence	$PSAL_t = PSAL_{t+1}, 800 < PRES_t$	1.0
f_4	Depth error	$PRES_t <= PRES_{t-1}$	1.0
f_5	Negative density inversion	$f_{SVM}(PDEN_t - PDEN_{MAX}, PRES_t) = 4$	1.0
f_6	Positive density inversion	$f_{SVM}(PDEN_{MIN} - PDEN_t, PRES_t) = 4$	1.0
f_7	Offset	$\lvert PSAL_t - \overline{PSAL}_t^{(WOA)} \rvert > 3\sigma(PSAL_t^{(WOA)})$ in 70% of all observed layers	1.0
f_8	Offset	$\lvert PSAL_t - \overline{PSAL}_t^{(past)} \rvert > 0.05$ in 70% of all observed layers	1.0
f_9	Offset	All salinity QC labels were 4 (bad) in a past profile obsevered by the same float	0.1

TABLE 5. COMPARISON ON QC LABEL ASSIGNMENT ACCURACY (ALL TYPES OF ERRORS).

Methods	TP	FN	TN	FP	Precision	Recall	F-value
Real-time QC (RQC)	10,314	8,201	128,243	426	0.961	0.557	0.705
Point-wise method with manually-designed features $\mathrm{PW_{man}}$	1,170	17,345	128,010	659	0.640	0.063	0.115
Point-wise method with automatically-desigend features $\mathrm{PW_{DT}}$	15,760	2,755	109,543	19,126	0.452	0.851	0.590
CRF with manually-designed features [9] $\mathrm{CRF_{man}}$	1,151	17,364	128,053	616	0.651	0.062	0.113
CRF with automatically-designed features (proposed) $\mathrm{CRF_{DT}}$	18,235	280	127,547	1,122	0.942	0.985	0.962

TABLE 6. COMPARISON ON QC LABEL ASSIGNMENT ACCURACY (FOUR TYPES OF ERRORS MARKED WITH '*' INT TABLE 1).

Methods	TP	FN	TN	FP	Precision	recall	F-value
Real-time QC (RQC)	169	140	35,725	303	0.358	0.547	0.433
Point-wise method with manually-designed features $\mathrm{PW_{man}}$	273	9	36,326	49	0.848	0.968	0.904
Point-wise method with automatically-desigend features $\mathrm{PW_{DT}}$	232	50	36,345	30	0.885	0.823	0.853
CRF with manually-designed features [9] $\mathrm{CRF_{man}}$	256	21	36,343	32	0.889	0.924	0.909
CRF with automatically-designed features (proposed) $\mathrm{CRF_{DT}}$	228	54	36,218	157	0.592	0.809	0.684

[3] S. Hosoda, "Argo float — innovation for autonomous observations of global oceans," in 9th Japanese-German Frontiers of Science Symposium, 2012, pp. 37–38.

[4] S. Levitus, J. Antonov, T. Boyer, O. Baranova, H. Garcia, R. Locarnini, A. Mishonov, J. Reagan, D. Seidov, E. Yarosh et al., "World ocean heat content and thermosteric sea level change (0-2000 m), 1955–2010," Geophys. Res. Lett., vol. 39, no. 10, 2012.

[5] S. Hosoda, T. Suga, N. Shikama, and K. Mizuno, "Global surface layer salinity change detected by argo and its implication for hydrological cycle intensification," Journal of Oceanography, vol. 65, no. 4, pp. 579–586, 2009. [Online]. Available: http://dx.doi.org/10.1007/s10872-009-0049-1

[6] T. Kagimoto, Y. Miyazawa, X. Guo, and H. Kawajiri, "High resolution kuroshio forecast system: description and its applications," in High Resolution Numerical Modelling of the Atmosphere and Ocean. Springer, 2008, pp. 209–234.

[7] e. a. TOZUKA, Tomoki, "Decadal modulations of the indian ocean dipole in the sintex-f1 coupled gcm," Journal of climate, pp. 2881–2894, 2007.

[8] S. Ono, H. Matsuyama, K.-i. Fukui, and S. Hosoda, "A preliminary study on quality control of oceanic observation data by machine learning methods," in Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems, Volume 1. Springer, 2015, pp. 679–693.

[9] S. Ono, H. Matsuyama, K. Fukui, and S. Hosoda, "Error detection of oceanic observation data using sequential labeling," in Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on, Oct 2015.

[10] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proceedings of the Eighteenth International Conference on Machine Learning, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. [Online]. Available: http://dl.acm.org/citation.cfm?id=645530.655813

[11] K. Seymore, A. McCallum, and R. Rosenfeld, "Learning hidden markov model structure for information extraction," in AAAI-99 Workshop on Machine Learning for Information Extraction, 1999, pp. 37–42.

[12] L. Olshen, C. J. Stone et al., "Classification and regression trees," Wadsworth International Group, vol. 93, no. 99, p. 101, 1984.