Weighted Ensemble of Gross Error Detection methods based on Particle Swarm Optimization

Daniel Dobos, Tien Thanh Nguyen, John McCall, Allan Wilson, Phil Stockton, Helen Corbett
School of Computing, Robert Gordon University, United Kingdom
Accord Company, Aberdeen, United Kingdom
{d.dobos,t.nguyen11,j.mccall}@rgu.ac.uk;{allan.wilson,phil.stockton,helen.corbett}@Accord-ESL.com

ABSTRACT

Gross errors, a kind of non-random error caused by process disturbances or leaks, can make reconciled estimates can be very inaccurate and even infeasible. Detecting gross errors thus prevents financial loss from incorrectly accounting and also identifies potential environmental consequences because of leaking. In this study, we develop an ensemble of gross error detection (GED) methods to improve the effectiveness of the gross error identification on measurement data. We propose a weighted combining method on the outputs of all constituent GED methods and then compare the combined result to a threshold to conclude about the presence of the gross error. We generate a set of measurements with or without gross error and then minimize the GED error rate of the proposed ensemble on this set with respect to the combining weights and threshold. The Particle Swarm Optimization method is used to solve this optimization problem. Experiments conducted on a simulated system show that our ensemble is better than all constituent GED methods and two ensemble methods.

CCS CONCEPTS

• **Computing methodologies** → *Ensemble methods*; • **Mathematics of computing** → *Evolutionary algorithms*;

KEYWORDS

Gross Error Detection, Ensemble Method, Particle Swarm Optimization, Ensemble Learning

ACM Reference Format:
Daniel Dobos, Tien Thanh Nguyen, John McCall, Allan Wilson, Phil Stockton, Helen Corbett. 2021. Weighted Ensemble of Gross Error Detection methods based on Particle Swarm Optimization. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3449726.3459415

1 INTRODUCTION

Let us consider a hydrocarbon process plant including four production wells and nine streams in Fig 1. It is required that the measurements for total input must be equal to the measurement of total output e.g. S1-V=S2-V+S2-L. However, due to measurement noise, i.e. random errors, this balance is somehow violated. Data reconciliation is a technique that corrects measurement errors to ensure several required constraints in the system [2]. It is noted that the reconciled estimates can be very inaccurate and even infeasible if non-random errors (called gross errors) are present. Thus, it is crucial to identify gross errors in measurement before applying a data reconciliation technique [3]. A GED method aims to detect whether the gross error presents in a measurement. In general, the detection is done by using statistical hypothesis tests [3]. In this study, we construct an ensemble of GED methods to further improve the effectiveness of GED. The outputs of GED methods are combined by using a weighted combining method to obtain the collaborated output. This output is then compared to a threshold to determine the presence of gross error. We propose to search for the combining weights and the threshold by minimizing the error rate of the GED task. The search process is conducted using Particle Swarm Optimisation (PSO), an effective computational intelligence method for heuristic searching [1][6].

GECCO '21 Companion, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3459415

2 PROPOSED METHOD

2.1 Ensemble of Detection

We denote $\mathbf{x}_n$ as a vector of measurements and H_i as a GED method that tests the null hypothesis that no gross error is on $\mathbf{x}_n$. In fact H_i works on $\mathbf{x}_n$ and outputs a probability $P_i(\mathbf{x}_n)$ (called p-value) of obtaining the observed results assuming that the null hypothesis is correct. By comparing $P_i(\mathbf{x}_n)$ to a chosen significance threshold α, we can come up with the rejection of the null hypothesis if $P_i(\mathbf{x}_n) < \alpha$ or a further consideration if $P_i(\mathbf{x}_n) \geq \alpha$.

In this study, we design an ensemble of K GED methods $\{H_i\}$ $i = 1, .., K$ in which the p-values of constituent methods are combined by a combining method C to obtain the collaborated decision $C\{P_i(\mathbf{x}_n)\}$ $i = 1, .., K$. In the ensemble method, the results of some methods are compensated by those of the other ones, which makes the collaborated result better than that of each constituent method. In [4], an ensemble of GED methods was introduced by using the Fisher method to combine the p-values. It is noted that the Fisher method requires all constituent methods independent. This assumption makes this ensemble difficult in choosing the constituent GED methods because some of them are related [3].

Normally, all methods are treated equally when combining in the ensemble e.g. all methods are associated with equal weights. This however may downgrade ensemble performance [5]. Here we propose associate methods $\{H_i\}$ $i = 1, .., K$ with the weights $\{w_i | 0 \leq w \leq 1\}$ $i = 1, ..., K$ and their p-values are combined based on these weights. The weighted combining of $P_i(\mathbf{x}_n)$ on $\mathbf{x}_n$ is given by:

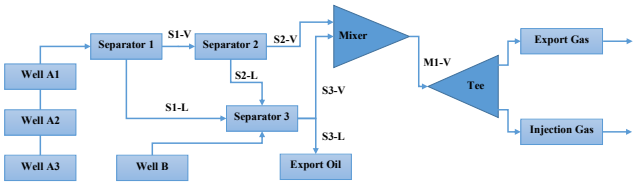

Figure 1: The hydrocarbon process plant in the experiment.

$$P(\mathbf{x}_n) = \sum_{i=1}^{K} w_i P_i(\mathbf{x}_n) \tag{1}$$

The combined value $P_i(\mathbf{x}_n)$ will be compared to a threshold to obtain the conclusion concerning the presence of gross error.

$$\begin{cases} P(\mathbf{x}_n) < \alpha & \text{Gross error is presented} \\ P(\mathbf{x}_n) \geq \alpha & \text{otherwise} \end{cases} \tag{2}$$

One question that arises from this model is to search for suitable weights w_i and also the threshold α for a particular system.

2.2 Optimisation Approach

We model an optimisation problem to search for the optimal value of combining weights and threshold. We first generate a set of N measurements (called training set) with or without a gross error (called ground truth $\mathbf{y}_n \in \{0, 1\}$). We then run each of $\{H_i\}$ $(i = 1, ..., K)$ on each measurement $\mathbf{x}_n$ $(n = 1, ..., N)$ in the training set to obtain $P_i(\mathbf{x}_n)$. For a set of weights $\{w_i | 0 \leq w \leq 1\}$, $P_i(\mathbf{x}_n)$ are combined by using (1) before using (2) to determine whether the gross error is presented. Here we minimize the 0-1 loss function on the training data to find the optimal value for w_i and α by comparing the detection result obtained from (2) to the ground truth of measurements. The optimisation problem is given by:

$$\min_{\{w_i\},\alpha} \left\{ 1 - \frac{1}{N} \sum_{n=1}^{N} \left(\mathbb{I}[\mathbb{I}[P(\mathbf{x}_n) < \alpha] = \mathbf{y}_n] \right) \right\} \quad \text{s.t.} \{w_i\}, \alpha \in [0, 1]; i = 1, ..., K \tag{3}$$

in which $\mathbb{I}[.] = 1$ if the condition is true, otherwise equal 0, $\mathbf{y}_n$ is the ground truth of $\mathbf{x}_n$. In this study, we use PSO [1] to solve the problem in (3). PSO is a popular swarm intelligence method that searches for the optimal value by iteratively trying to improve a candidate solution concerning a given measure of quality. The position of each particle encodes w_i and α while the fitness is calculated on the training set with the value of each candidate w_i and α. The optimal value for w_i and α will be obtained after a number of iterations.

3 EXPERIMENTAL STUDIES

3.1 Dataset and Settings

We conducted the experiments on the plant illustrated in Fig.1. We used CHARM simulation package [4] which outputs a vector for measurements in which no gross errors are present. We randomly added gross errors to this vector by changing the magnitude of any one of the six streams by +5% and +25% [4]. From this data, we generated a training set with 1600 observations; half of it has the gross error. We also generated two test sets containing 7400 and 1800 observations for the evaluation [4]. To construct the ensemble, we used five GED methods namely Global Test, Principle Component Analysis Test (PCA), Measurement Statistic Test (MST), Constraint Test, and Generalized Likelihood Ratio Test (GLR) [3]. The proposed ensemble was compared to the Fisher method-based ensemble system namely Ensemble (Fisher) and Ensemble (Selection) introduced in [4]. In the PSO algorithm, we used the settings in [4]. The accuracy and the F1 score of all methods are shown in Table 1.

Table 1: The experimental results

Methods	Test Case 1		Test Case 2	
	Accuracy	F1 Score	Accuracy	F1 Score
Global Test	0.2081	0.1887	0.2683	0.2666
PCA	0.1832	0.1685	0.2494	0.2480
MST	0.4737	0.3545	0.5017	0.4446
Constraint Test	0.3277	0.2713	0.3733	0.3547
GLR	0.4441	0.3390	0.4772	0.4290
Ensemble (Fisher)	0.5253	0.3786	0.5422	0.4650
Evolved Ensemble (Selection)	0.5873	0.4073	0.5850	0.4875
Proposed Ensemble	**0.6031**	**0.4141**	**0.6039**	**0.4979**

3.2 Result and Discussions

The proposed ensemble is better than all constituent methods on both test sets. In detail, our method is 39.5%, 41.99%, 12.94%, 27.54%, and 15.9% better than Global Test, PCA, MST, Constraint Test, and GLR for accuracy on the test set 1. For the F1 score, the proposed method is also remarkably better than all methods, for example, 12.94% better than MST, the best constituent method in our experiment. Meanwhile, our ensemble is about 7.78% and 6.17% better than the Ensemble (Fisher) for accuracy in test set 1 and 2, respectively. Our ensemble is also marginally better than Ensemble (Selection) (0.4141 vs. 0.4073 and 0.4979 vs. 0.4875 for F1 Score on the test set 1 and 2, respectively. The experimental results show the outperformance of our weighted combining approach compared to not only the five constituent methods but also the Fisher method-based ensemble and ensemble with selection. The optimal value of the weights is (0.8954, 0.0029, 0.7186, 0.0340, 0.5875) for Global Test, PCA, MST, Constraint Test, and GLR Test. It is a surprising result since the Global Test is not the best method but its associated weight is highest among all the weights. PCA is the poorest GED method and it contributes very small to the collaborated result since its weight is nearly equal to 0.

REFERENCES

[1] James Kennedy and Russell Eberhart. 1995. Particle swarm optimization. In *Proceedings of ICNN'95-International Conference on Neural Networks*, Vol. 4. IEEE, 1942–1948.
[2] José Loyola-Fuentes and Robin Smith. 2019. Data reconciliation and gross error detection in crude oil pre-heat trains undergoing shell-side and tube-side fouling deposition. *Energy* 183 (2019), 368–384.
[3] Shankar Narasimhan and Cornelius Jordache. 1999. *Data reconciliation and gross error detection: An intelligent use of process data.* Elsevier.
[4] Tien Thanh Nguyen et al. 2020. Evolved ensemble of detectors for gross error detection. *GECCO Companion 2020* (2020), 281–282.
[5] Tien Thanh Nguyen et al. 2020. Evolving interval-based representation for multiple classifier fusion. *Knowledge-Based Systems* (2020).
[6] RE Perez and K Behdinan. 2007. Particle swarm approach for structural design optimization. *Computers & Structures* 85, 19-20 (2007), 1579–1588.