Monitoring of Drinking-water Quality by Means of a Multi-objective Ensemble Learning Approach

Victor Henrique Alves Ribeiro
Industrial and Systems Engineering Graduate Program
Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba, Paraná, Brazil
victor.henrique@pucpr.edu.br

Gilberto Reynoso-Meza
Industrial and Systems Engineering Graduate Program
Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba, Paraná, Brazil
g.reynosomeza@pucpr.br

ABSTRACT

This paper proposes the use of multi-objective ensemble learning to monitor drinking-water quality. Such problem consists of a data set with an extreme imbalance ratio where the events, the minority class, must be correctly detected given a time series denoting water quality and operative data on a minutely basis. First, the given data set is preprocessed for imputing missing data, adjusting concept drift and adding new statistical features, such as moving average, moving standard deviation, moving maximum and moving minimum. Next, two ensemble learning techniques are used, namely SMOTEBoost and RUSBoost. Such techniques have been developed specifically for dealing with imbalanced data, where the base learners are trained by adjusting the ratio between the classes. The first algorithm focuses on oversampling the minority class, while the second focuses on under-sampling the majority class. Finally, multi-objective optimisation is used for pruning the base models of such ensembles in order to maximise the prediction score without reducing generalisation performance. In the training phase, the model is trained, optimised and evaluated using hold-out validation on a given training data set. At the end, the trained model is inserted into a framework, which is used for online event detection and assessing the model's performance.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Ensemble methods*; *Continuous space search*;

KEYWORDS

Machine learning, evolutionary computation, time series, anomaly detection.

ACM Reference format:
Victor Henrique Alves Ribeiro and Gilberto Reynoso-Meza. 2019. Monitoring of Drinking-water Quality by Means of a Multi-objective Ensemble Learning Approach. In *Proceedings of Genetic and Evolutionary Computation Conference Companion, Prague, Czech Republic, July 13–17, 2019 (GECCO '19 Companion),* 2 pages.
https://doi.org/10.1145/3319619.3326745

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07.
https://doi.org/10.1145/3319619.3326745

1 INTRODUCTION

This paper proposes the application of multi-objective ensemble learning for solving the "GECCO 2019 Industrial Challenge: Monitoring of drinking-water quality" [3], based on a previous competition entry [6]. The challenge is composed of a classification problem where a time series with six water quality data and operational data features are given in order to detect events. Such events represent the minority class on an extremely imbalanced scenario.

In order to solve such problem, the use of feature engineering, ensemble learning techniques and evolutionary multi-objective optimisation is proposed. Specifically, the proposed solutions makes use of one of two different approaches for imbalanced ensemble learning: boosting with synthetic minority oversampling technique (SMOTEBoost) [1]; and boosting with random undersampling (RUSBoost) [7]. In order to maximise the predictions of the ensemble without loosing generalisation performance, the multi-objective evolutionary algorithm based on decomposition (MOEA/D) [8] is used for pruning the ensemble's base models. Later, in order to select one of the non-dominated solutions, physical programming [2] is used.

The remainder of this document is organised as follows: Section 2 details the proposed techniques used for data preprocessing, ensemble learning and multi-objective optimisation; while Section 3 concludes the paper.

2 PROPOSAL

2.1 Preprocessing

First, in order to adjust the data set, missing points are imputed using the average value of the past 30 minutes. Next, in order to reduce the effects of concept drift, linear detrending is performed on all six features using data from the past 24 hours. The detrended signals are used as additional features.

The next step focuses on creating new features for classification. Since the problem deals with time series, signal processing and statistical techniques are used in order to create the following new features: differences between the current and previous samples, moving average, moving standard deviation, moving maximum and moving minimum. All moving methods are applied on a window of 30 minutes.

Finally, for training and assessing the performance of the trained models, the data set is split for hold-out validation in the following manner: 60% of the data is used for training the ensemble, 20% is used for evaluating the performance during optimisation, and the remaining 20% is used for evaluating the model selected with physical programming.

Table 1: Preference matrix for model selection. Five preference ranges have been defined: highly desirable (HD), desirable (D), tolerable (T) undesirable (U) and highly undesirable (HU).

Preference Matrix					
	$\leftarrow$ HD	$\rightarrow \leftarrow$ D	$\rightarrow \leftarrow$ T	$\rightarrow \leftarrow$ U	$\rightarrow \leftarrow$ HU $\rightarrow$
Objective	J_i^0	J_i^1	J_i^2	J_i^3	J_i^4 J_i^5
J_1	0.00	$0.25 \cdot J_1(\boldsymbol{x_d})$	$0.50 \cdot J_1(\boldsymbol{x_d})$	$J_1(\boldsymbol{x_d})$	$1.5 \cdot J_1(\boldsymbol{x_d})$ 1.00
J_2	0.00	$0.25 \cdot J_2(\boldsymbol{x_d})$	$0.50 \cdot J_2(\boldsymbol{x_d})$	$J_2(\boldsymbol{x_d})$	$1.5 \cdot J_2(\boldsymbol{x_d})$ 1.00
J_3	1	$0.50 \cdot M$	$0.75 \cdot M$	M	- -

2.2 Imbalanced Ensemble Learning

Since the problem in focus is composed of an imbalanced data set, two ensemble learning techniques designed for such scenarios are used, the SMOTEBoost and RUSBoost. In both techniques, the data set is balanced before training the base learners. While the first technique performs a synthetic oversampling of the minority class, the second performs the random under-sampling of the majority class.

2.3 Multi-Objective Ensemble Pruning

In order to optimise the performance of the previous ensembles, but without loosing generalisation power, a multi-objective optimisation design (MOOD) procedure [5] is performed. First, a multi-objective problem (MOP) is formulated to prune the ensemble's base models and tune the decision threshold. Next, such problem is optimised using MOEA/D, a multi-objective optimisation (MOO) algorithm, which returns a set of non-dominated ensembles. Finally, in the multicriteria decision making (MCDM) stage, physical programming is used to select a final preferred model.

The formulation of the MOP is performed as follows:

$$\min_{\boldsymbol{x}} \boldsymbol{J(x)} = [J_1(\boldsymbol{x}), J_2(\boldsymbol{x}), J_3(\boldsymbol{x})] \tag{1}$$

subject to:

$$\boldsymbol{x} = [\boldsymbol{m}, t] \tag{2}$$

$$m_i \in \{0, 1\}, i = [1, \ldots, M] \tag{3}$$

$$0.0 \leq t \leq 1.0 \tag{4}$$

where the objectives are: false positive rate ($J_1(\boldsymbol{x})$); false negative rate ($J_2(\boldsymbol{x})$); and the ensemble's complexity ($J_3(\boldsymbol{x})$), defined as the number of selected base models. The decision variables are: the binary selection of each of the M base models (m_i); and the decision threshold for the ensemble's prediction (t).

A set of non-dominated ensembles is generated by optimising the aforementioned MOP with the MOEA/D algorithm. After such step, the final model is found by selecting the model with best physical programming ranking [2] according to the preference matrix in Table 1 and the following equations:

$$\boldsymbol{x_d} = [\boldsymbol{m_d}, t_d] \tag{5}$$

$$m_{di} = 1, i = [1, \ldots, M] \tag{6}$$

$$t_d = 0.5 \tag{7}$$

indicating that $\boldsymbol{x_d}$ is a decision vector where the selection of each of the base learners m_{di} is set to 1, and the decision threshold t_d is set to 0.5. Such values indicate that a tolerable solution presents, at least, the same predictive performance as an ensemble without pruning. On the one hand, desirable and highly desirable solutions presents, respectively, 50% and 75% improvement in predictive performance and 25% and 50% reduction of the total number of base models. On the other hand, the reduction in predictive performance indicates undesirable and highly undesirable solutions. Such approach is based on the procedure found in [4].

3 CONCLUSION

The present document proposes a multi-objective ensemble learning approach for creating an online drinking-water quality monitoring system. To do so, MOO is applied to prune base learners from two different imbalanced ensemble learning algorithms in order to build a model with high predictive performance and generalisation. Such procedure can be used as a base solution for an online drinking-water event detector on a real-world system.

ACKNOWLEDGMENTS

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) and the *Conselho Nacional de Pesquisa e Desenvolvimento* (CNPq) - Brazil - Finance Codes: 159063/2017-0-PROSUC, 304066/2016-8-PQ2, 437105/2018-0-Univ.

REFERENCES

[1] Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. 2003. SMOTEBoost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*. Springer, 107–119.

[2] Achille Messac. 1996. Physical programming: effective optimization for computational design. *AIAA journal* 34, 1 (1996), 149–158.

[3] Frederik Rehbach, Steffen Moritz, and Thomas Bartz-Beielstein. 2019. GECCO 2019 Industrial Challenge: Monitoring of drinking-water quality. (2019). https://www.th-koeln.de/mam/downloads/deutsch/hochschule/fakultaeten/informatik_und_ingenieurwissenschaften/rulesgeccoic2019.pdf

[4] Gilberto Reynoso-Meza, Javier Sanchis, Xavier Blasco, and Roberto Z Freire. 2016. Evolutionary multi-objective optimisation with preferences for multivariable PI controller tuning. *Expert Systems with Applications* 51 (2016), 120–133.

[5] Gilberto Reynoso-Meza, Javier Sanchis, Xavier Blasco, and Miguel Martínez. 2016. Preference driven multi-objective optimization design procedure for industrial controller tuning. *Information Sciences* 339 (2016), 108–131.

[6] Victor Henrique Alves Ribeiro and Gilberto Reynoso-Meza. 2018. Online anomaly detection for drinking water quality using a multi-objective machine learning approach. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. ACM, 1–2.

[7] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2010. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40, 1 (2010), 185–197.

[8] Qingfu Zhang and Hui Li. 2007. MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on evolutionary computation* 11, 6 (2007), 712–731.