Pre-processing of imbalanced samples and the effective contribution in fault diagnosis in wastewater treatment plants

Y. G. Xu, W. K. Deng, B. Song, X. Y. Deng and F. Luo

ABSTRACT

Fault diagnosis by machine learning techniques is of great importance in wastewater treatment plants (WWTPs). A key factor influencing the accuracy of fault diagnosis lies in the imbalance between the sample data in minority classes (i.e. faulty situations) and that in majority classes (i.e. normal situations), which may cause misjudgments of faults and lead to failure in practical use. This study proposes a novel pre-processing method with a fast relevance vector machine (Fast RVM) reducing the data of majority class samples and the synthetic minority over-sampling technique expanding the minority class samples. A case study indicates that this pre-processing method could be a promising solution for imbalanced data classification in WWTPs and the pre-processed data can be well diagnosed by back-propagation neural networks, support vector machine, RVM and Fast RVM models.

Key words | data classification, imbalanced data, pre-processing, wastewater treatment plant

Y. G. Xu (corresponding author)
W. K. Deng
X. Y. Deng
F. Luo
School of Automation Science and Engineering,
South China University of Technology,
Guangzhou 510640,
China
E-mail: *xuyuge@scut.edu.cn*

B. Song
School of Chemical and Petroleum Engineering,
Curtin University,
GPO Box U1987,
Perth,
Western Australia 6845,
Australia

INTRODUCTION

Monitoring and correcting the possible operation faults during wastewater treatment is of great importance, despite the fact that the process can be affected by various factors such as complex biological reaction mechanisms, highly time-varying and multivariable aspects (Hong *et al.* 2003). An optimal real-time fault diagnosis method can help maintain the stability of the water processing flow in wastewater treatment plants (WWTPs) and benefit the industry with adequate water output, lower operation cost and less secondary pollution to the environment (Hu & Hu 2002; Hamed *et al.* 2004). However, due to the imbalance of data numbers between the situation under faulty states and ground states, misclassification and inaccurate diagnosis during water processing often occurs and leads to inadequate output and even damage to the equipment as the traditional classifiers bias to the majority classes due to more sample data compared with that of the minority. The classification between imbalanced classes is considered as the main challenge of fault diagnosis using machine learning techniques.

Research on classifying imbalanced data began in the 1960s (Cover & Hart 1967). One way to improve the classification is to improve the algorithm of traditional machines that are used for balanced data classification or develop new algorithms because the previous machines assume or expect balanced class distributions or equal misclassification costs (He & Garcia 2009). For instance, the improvement of support vector machines (SVM) (Raskutti & Kowalczyk 2004; Hwang *et al.* 2011), and the development of back-propagation neural networks (BPNNs) (Chen *et al.* 2008) and extreme learning machines (Zong *et al.* 2013) are effective ways to compensate or increase the weight of data from minority classes. However, the weight of data strongly depends on the distribution of given datasets and needs to be adjusted manually according to different raw data because the imbalance of data numbers in different classes still exists in these

doi: 10.2166/hydro.2017.206

methods. Another way is to minimize the gap between majority and minority classes through data pre-processing by either under-sampling the datasets of majority classes (Polat & Güneş 2008; Yen & Lee 2009) or over-sampling datasets from minority classes (Chawla *et al.* 2002; Nguyen *et al.* 2011). Data pre-processing has been recognized as an effective method for imbalanced data classification. However, the presented methods for data under-sampling have a risk of losing the representativeness of datasets, suggesting that more study is needed to understand how to minimize the loss of representative datasets during under-sampling. In the industry of wastewater treatment, unfortunately most of the previous studies use ideally equivalent datasets for fault diagnosis (Fuente & Vega 1995; Lee *et al.* 2005; Motamarri & Boccelli 2012; Tao *et al.* 2013) and neglect the imbalance of practical data. It should be noted that a previous work by Qian *et al.* (2014) has enhanced the accuracy of minority data classification to some extent, which implies that data pre-processing can also be an applicable route to increasing the accuracy of fault diagnosis in WWTPs.

Based on the existing studies regarding data pre-processing in other areas and the present status of fault diagnosis in wastewater treatment, this study proposes a novel method for imbalanced data classification and applies it in a case study of wastewater treatment, where the synthetic minority over-sampling technique (SMOTE) algorithm is used for over-sampling and Fast RVM for under-sampling. The fault diagnosis performance with different classifiers is shown to be enhanced compared with the results without pre-processing.

DATA AND METHODOLOGY

Introduction of the raw data

The raw data in this study are obtained from the UCI (University of California Irvine 1987) Machine Learning Repository (http://archive.ics.uci.edu/ml/index.html), which is primarily obtained from the daily measurements in an urban WWTP in Manresa, a town located near Barcelona, Spain. The single plant treats a daily wastewater flow of some 35,000 m^3 with pretreatment, primary treatment by clarification, secondary treatment by means of activated sludge, and finally chlorination. The same raw data has been used for wastewater plant data processing elsewhere (Sànchez *et al.* 1997; Fan 2003; Zhang & Zhu 2011; Liu & Han 2013; Qian 2014). In total, 527 samples with 38 attributes were collected daily over approximately 2 years. Definitions of the 38 attributes are shown in Appendix A. The 527 samples are classified into 13 classes named by an Arabic numeral, as shown in Appendix B. (Appendix A and Appendix B are available with the online version of this paper.) Each class represents one type of wastewater treatment operation. The distribution of raw dataset numbers under the 13 classes is shown in Table 1. The data show that at faulty situations (i.e. class 2, 3, 4, 6, 7, 8, 10), the number of datasets is limited to no more than four, compared with normal situations (i.e. class 1, 5, 9 and 11) at more than 53. Based on this dramatic imbalance in data between faulty and normal situations, more sampled datasets will tend to represent the normal situations and cause inaccurate diagnosis of classifiers without pre-processing.

Methodology

Summary of methodology

The framework of the method is shown in Figure 1. As shown, the data from the majority and minority classes are pre-processed separately to neutralize the imbalance before further diagnosis. The majority class data are reduced via *K*-means clustering and the fast relevance vector machine (Fast RVM) algorithm (Tipping & Faul 2003), and the minority are expanded based on the SMOTE algorithm (Chawla *et al.* 2002). Continuously, the data are combined together for classification study with different classifiers.

Table 1 | Distribution of raw dataset numbers under 13 classes

Class	1	2	3	4	5	6	7	8	9	10	11	12	13
Number of instances	279	1	1	4	116	3	1	1	65	1	53	1	1

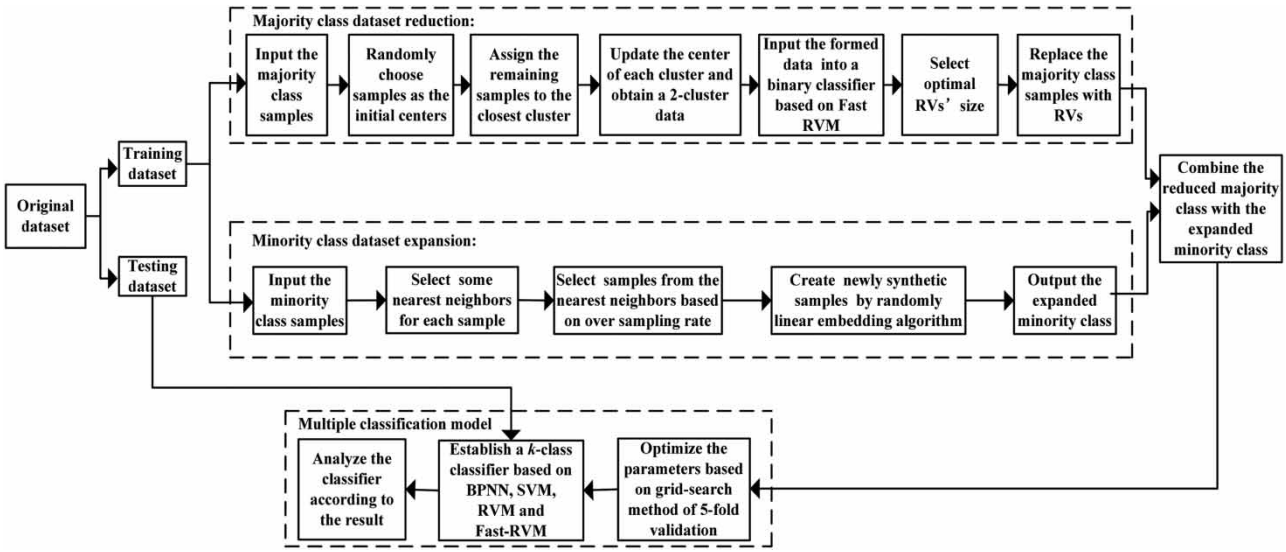

Figure 1 | Sketch for the proposed data processing and fault diagnosis method.

Specifically, one multiple classifier BPNN, and three integrated multiple classifiers based on SVM, RVM and Fast RVM are each used for fault diagnosis to evaluate the effect of this pre-processing method onto different classifiers.

Main algorithms used

Method for under-sampling of the majority class data. Considering the possible loss of useful data caused by the reduction of datasets in the majority classes, it is important to choose a proper model that can reduce the number of datasets and maintain the maximum representativeness of the remaining data. This study employed Fast RVM rather than SVM and RVM for the following reasons.

Generally, all of the three methods share common steps to reduce the datasets, as shown in Figure 1. The main difference lies in choosing representative datasets (i.e. support vectors (SVs) for SVM and relevant vectors (RVs) for RVM & Fast RVM). For SVM, the machine learns and updates the data through minimizing the gap between the samples and hyperplane based on the Structural Risk Minimization Principle. However, the sparsity of data could become weaker with the increase of SVs, if the training datasets had been increased (Burges 1996; Burges & Scholkopf 1997). The RVM is a sparse Bayesian learning framework for producing sparse models, in which a set of hyperparameters is placed based on the principle of automatic relevant data points. After training, the weights of most samples tend to zero and only a few samples with non-zero weights will be left as RVs. Thus for a certain class of datasets, the number of RVs for RVM should be less than the number of SVs for SVM, which suggests that RVM can collect fewer samples with good representability after data reduction (Tipping 2001; Yang 2006), even though both methods can reduce the loss of representative datasets. Based on RVM, Fast RVM keeps the sparsity of RVs and reduces the computational complexity that can reduce the time complexity of training of a RVM classifier (Tipping & Faul 2003). Thus Fast RVM is chosen as a fast and proper model for majority class data reduction.

In data reduction, the majority classes are first divided into two clusters by *K*-means algorithm (Hartigan & Wong 1979). Then, the data from two clusters are input to a Fast RVM to build up a binary classifier. RVs are obtained from the established Fast RVM model. The number of RVs is fewer than the original database, but is representative of the sparsity of the original data, and plays a key role in generalizability (Liu 2011). Thus, we replace the majority class data with these RVs. In this way, the majority class is reduced and the effective information is retained. The reduction procedures are as follows:

(1) cluster the majority class samples with K-means algorithm;
(2) build up a binary classifier on Fast RVM;
(3) replace the majority class samples with RVs.

The steps of Fast RVM classification are introduced as follows (Tipping & Faul 2003):

The clustered majority class is defined as input-target pairs $\{z_i, t_i\}_{i=1}^{N}$, $z_i \in R^d$, $t_i \in R$, where N is the number of the dataset, i is the numbering of samples, d is the number of attributes. The predictive function is as shown in Equation (1):

$$t_i = y(z_i; w) + \varepsilon_i \tag{1}$$

where function $y(z)$ is defined as Equation (2):

$$y(z; w) = \sum_{i=1}^{N} w_i K(z, z_i) + w_0 \tag{2}$$

where $K(z, z_i)$ is a kernel function; w_i is the adjustable parameters (or 'weights') of the basis function, $w = [w_0, w_1, \ldots, w_N]^{\mathrm{T}}$, $\varepsilon_i \sim N(0, \sigma^2)$. Thus, $t_i \sim N(y(z_i, w), \sigma^2)$. Due to the assumption of independence of the predictive function t_i, it is easy to obtain the likelihood of the complete data $p(t|\sigma^2,w)$.

To avoid over-fitting, it is assumed that the weights are modeled probabilistically as independent zero-Gaussian with variance α_i^{-1}, where α is a vector of N+1 hyperparameters. So the corresponding target of a new test sample should be t^* in terms of the predictive distribution $p(t*|\mathrm{t}) \sim \mathrm{p}(\mathrm{w}, \alpha, \sigma^2|\mathrm{t})$. According to the prior distribution and the maximum-likelihood estimation distribution, the posterior can be decomposed as shown in Equation (3):

$$p\left(\frac{w, \alpha, \sigma^2}{t}\right) = p\left(\frac{w}{t, \alpha, \sigma^2}\right) p\left(\frac{\alpha, \sigma^2}{t}\right) \tag{3}$$

To approximate Equation (3), the hyperparameter posterior $p(\alpha, \sigma^2|t) \propto p(t|\alpha, \sigma^2)p(\alpha)p(\sigma^2)$ is represented by delta-function at its most-probable values α_{MP}, σ^2_{MP}. Herein, sparse Bayesian 'learning' is formulated as the (local) maximization with α representing the marginal likelihood, or equivalently, the logarithm of which is $L(\alpha) = log[p(t|\alpha, \sigma^2)]$, which is further transformed into Equation (4):

$$\begin{aligned} L(\alpha) &= L(\alpha_{-i}) + \frac{1}{2}\left[\log \alpha_i - \log(\alpha_i + S_i) + \frac{(Q_i)^2}{\alpha_i + S_i}\right] \\ &= L(\alpha_{-i}) + l(\alpha_i) \end{aligned} \tag{4}$$

where $L(\alpha_{-i})$ is the marginal likelihood with the contribution of basis vector ϕ_i removed when $\alpha_i = \infty$, and $l(\alpha_i)$ represents the isolated part in $L(\alpha)$ where terms in α_i. S_i is the 'sparsity factor' and Q_i is the 'quality factor'. More details about the definition and derivation of S_i and Q_i can be observed in Tipping & Faul (2003).

$L(\alpha)$ has a unique maximum according to the value of α_i:

$$\alpha_i = \begin{cases} \dfrac{S_i^2}{Q_i^2 - S_i} & Q_i^2 > S_i \\ \infty & Q_i^2 \le S_i \end{cases} \tag{5}$$

Based on Equation (5), an iterative analysis is followed to find the set of weights that maximize the function $L(\alpha)$, in which the hyperparameter α is associated with the corresponding updated weight w. After updating, the trained classifier $y(z; w) = \sum_{i=1}^{N} w_i K(z, z_i) + w_0$ is built up with samples that have weight w unequal to 0, and the corresponding samples are RVs.

Method for over-sampling of the minority class data. The minority classes are expanded based on the SMOTE algorithm because of the optimal results illustrated in earlier studies by Chawla *et al.* (2002), Blagus & Lusa (2013) and Gao *et al.* (2011). This method can generate more synthetic samples of the minority class in the new feature space based on the linear embedding algorithm (Equation (6)). The over-sampling process method is created as follows:

(1) Calculate the Euclidean distances from a minority class sample p to the remaining samples in the minority class and record the subscripts of the nearest five samples. The number of the nearest value is recorded as M.
(2) The number of added samples equals the expanding ratio D (less than M) and the added samples are recorded as $q_1, q_2, \ldots, q_D$.

(3) Create new synthetic samples s_j for the minority class by randomly embedding a linear algorithm between the original sample p and $q_j (j = 1, 2, \cdots, D)$ described as follows:

$$s_j = p + \text{rand}(0, 1) * (\text{q}_\text{j} - \text{p}), \text{j} = 1, 2, \cdots, \text{D} \quad (6)$$

where rand(0, 1) is a random number between 0 and 1.

Multiple classification models. To characterize the effect of pre-processing on imbalanced data classification and its applicability with different classifiers, a multiple classifier BPNN (Rumelhart *et al.* 1986) and three multiple binary classifiers (SVM, RVM and Fast RVM) are used to classify the pre-processed datasets. These classifiers are commonly used for classification in many practical fields. For the multiple binary classifier, several single binary classifiers are combined together through the 'One-Against-One' (OAO) algorithm method (Hsu & Lin 2002; Galar *et al.* 2011) to allow for classifying multiple classes. Specifically, for a multi-classification problem with k classes, $(k(k-1)/2)$ binary classifiers are employed, each of which processes two different data from two different classes. Classification will be based on the voting of data, where each binary classifier casts one vote for its preferred class, and the final result is the class with the most votes. For example, suppose the classification function $f_{ij}(x)$ is employed to distinguish class i from class j. If $f_{ij}(x) < 0$, the data x will be allocated to class i, then class i gets one vote. Otherwise, the sample x will be allocated to class j, then class j gets one vote. The final classification result goes to the class with the most votes. The basic structure of this algorithm is shown in Figure 2.

The modeling process of the Fast RVM based binary classifier is the same as mentioned above under 'Methodology', and the modeling process of SVM and RVM classifiers are introduced elsewhere (Vapnik 1995; Tipping 2001).

Performance index

The classification performance of multiclass problems with imbalanced data is usually evaluated by G-mean, which is calculated as shown in the following equations:

$$G - \text{mean} = \left(\prod_{i=1}^{k} R_i \right)^{\frac{1}{k}} \quad (7)$$

$$R_i = \frac{n_{ii}}{\sum_{j=1}^{k} n_{ij}} \quad (8)$$

where k is the number of classes, n_{ii} is the number of the ith class samples correctly classified as the ith class, $n_{ij} (i \neq j)$ is the number of the ith class samples incorrectly classified as the jth class. R_i represents the recall of data in class i, which indicates the accuracy of classification for each class; while the index G-mean is defined as the geometric mean of the recall for all classes to express the classification accuracy

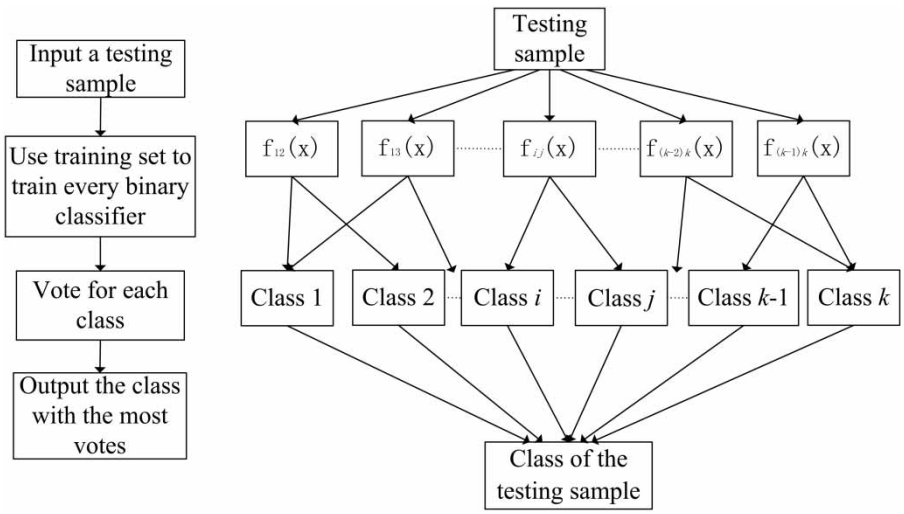

Figure 2 | 'OAO' algorithm for binary classifiers combination.

and balance between classes. Thus, G-mean value is selected as the value indicating the performance of classifiers using imbalanced data from wastewater processing data. Additionally, the accuracy of single class classification, total accuracy of classification and training time are also calculated and compared.

EXPERIMENTS, RESULTS AND DISCUSSION

Data pre-processing

According to the description of the 13 classes, the dataset is further classified into four classes (one faulty class and three normal classes) as shown in Tables 2 and 3. Similar classification methods have been conducted elsewhere (Fan 2003; Zhang & Zhu 2011; Liu & Han 2013; Qian 2014).

It is clear from Table 2 that the dataset is still imbalanced with the ratio of four classes at 23.7:8.3:4.6:1. It is worth noting that these wastewater datasets present an identical imbalanced distribution.

Initially, 147 data with incomplete attributes are eliminated. The remaining 380 original samples are defined as input-target pairs $\{x_i, t_i\}_{i=1}^{380}$, where $x_i \in R^{38}$, $t_i \in R$, i is the number of samples ($i = 1, 2, \ldots, 380$). For each of the 380 samples, the 38 attributes are normalized according to the formula $x_{normal}^{j} = \frac{x_i^j - x_{\min}^j}{x_{\max}^j - x_{\min}^j}$, where j is the numbering of attributes ($j = 1, 2, \ldots, 38$). For the training and testing, the 380 original samples are divided into training set $\{x_i, t_i\}_{i=1}^{n_1}$ and testing set $\{x_i, t_i\}_{i=1}^{n_2}$ at a ratio 2:1 by stratified random sampling, where n_1 and n_2 are the number of examples in training set and testing set respectively ($n_1 + n_2 = 380$).

Before classification, the training set $\{x_i, t_i\}_{i=1}^{n_1}$ is pre-processed by the proposed pre-processing method. Regarding the majority classes, class 1 is chosen as an identical majority class, the samples of which are reduced to approximately 20, 40, 60, 80, 100 and 120 RVs and comparisons are made to find an optimal input number for classification. Regarding the minority class 3 and 4, the data are expanded by the SMOTE method as introduced above under 'Performance index'. After pre-processing, the proposed Fast RVM based multiclass processing model is employed for testing sets $\{x_i, t_i\}_{i=1}^{n_2}$. The results are shown in Figures 3 and 4, where 'Total acc.' is the total classification accuracy rate, 'G-mean' represents the geometric mean of the recall for all classes, 'R1 acc.' indicates the accuracy rate of class 1, 'R2 acc.' indicates the accuracy rate of class 2 and so on.

As shown in Figure 3, when the number of RVs is approximately 80, the G-mean and total classification accuracy are both at their highest. In Figure 4, when the number of RVs is approximately 80, the classification accuracy of class 1 and class 4 are relatively higher than the classification accuracy of class 2 and class 3. The results suggest that the model performs well for class 1, when the number of RVs is at approximately 80. Distribution of the datasets is shown in Table 4.

Table 2 | Distribution of datasets in four classes

Class	1	2	3	4
Numbers	332	116	65	14

Table 3 | Descriptions of four classes

Class	Interpretation of classes
1	Normal situation
2	Normal situation with performance over the mean
3	Normal situation with low influent
4	Faults caused by the secondary settler problems, storm and solids overload

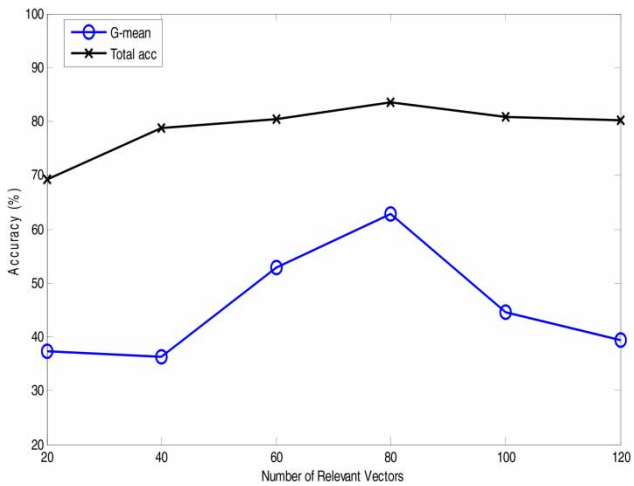

Figure 3 | G-mean and total classification accuracy of the model with different numbers of RVs.

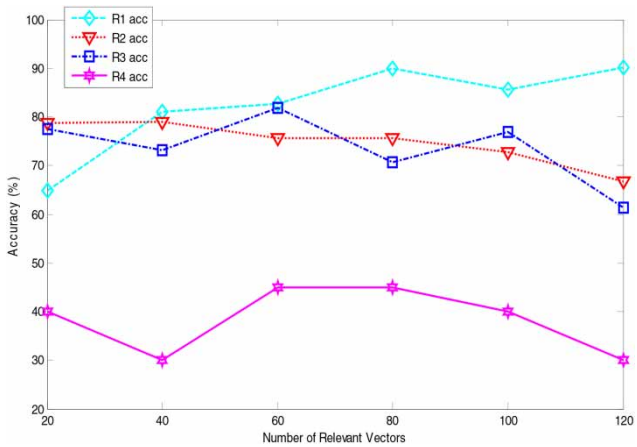

Figure 4 | Accuracy of each class for the model with different numbers of RVs.

From Table 4, it can be observed that the distribution of the original training datasets is 39.5:14.5:8:1. After pre-processing, the new distribution of the training datasets is 4:2.9:2.5:1, which suggests that the imbalance of datasets is minimized.

Simulation and analysis

Classification performance on original data

To analyze the classification performance on the pre-processed data, four classifiers, including a single multiple classifier BPNN, and three multiple binary classifiers based on SVM, RVM and Fast RVM, are chosen to classify the same data and the results are presented in Table 5. The single multiple classifier, the BPNN, is constructed with a three-layer structure, 38 input nodes and four output nodes, and the optimal hidden layer node number is defined by searching from 1 to 30 with a grid search mechanism on the training set. Regarding SVM, RVM and Fast RVM, radial basis function function $K(x, x_i) = \exp(-||x - x_i||^2/\sigma^2)$ is chosen as the kernel function, the penalty factor C in SVM and the kernel width σ in SVM, RVM and Fast RVM are determined from -10 to 10 through 5-fold cross-validation with a grid search mechanism on the training set. The OAO algorithm, a method combining several binary classifiers, is used to combine SVM, RVM and Fast RVM to yield a multiple classifier. The experimental classification performances are evaluated by the average values of 10 trials through 5-fold cross-validation. 'Time' in Table 5 indicates the training time. All experiments are simulated in the MATLAB 2013a simulation environment. The PC environment comprises: Core i3 2.1G, 2G ROM, more than 150 G remaining space of hard disk.

As shown in Table 5, majority classes using the BPNN classifier have high classification accuracies (i.e. class 1 accuracy R1 acc = 92.88%, class 2 accuracy R2 acc = 78.33%) because the accuracy of the BPNN relies heavily on the number of datasets. Therefore, the lack of samples in minority classes leads to poor classification accuracies, especially in class 4. Accuracy on class 4 is only 5%. The G-mean value of the BPNN classifier is 0.07. The results match the characteristics of the training principle of the BPNN, i.e. the empirical risk minimization principle, which strongly relies on the number of the training samples. Thus the imbalance between minority classes and majority classes leads to unsatisfactory results.

Table 4 | Distribution of datasets

	Class 1	Class 2	Class 3	Class 4	Total
Original datasets	238	88	48	6	380
Training sets	158	58	32	4	252
Pre-processed training sets	80 (±10%)	58	50	20	208
Testing sets	80	30	16	2	128

Table 5 | Classification performance of four classifiers with original data

Classifier	G-mean	R1 acc (%)	R2 acc (%)	R3 acc (%)	R4 acc (%)	Total acc (%)	Time (s)
BPNN	0.07	92.88	78.33	60.00	5.00	83.98	2.32
SVM	0.39	94.38	68.00	58.75	30.00	82.73	9.36
RVM	0.42	91.88	70.67	65.63	40.00	82.81	2.69
Fast RVM	0.29	93.47	71.48	60.42	22.22	83.07	1.45

Table 6 | Classification performances of four classifiers with pre-processed data

Classifier	G-mean	R1 acc (%)	R2 acc (%)	R3 acc (%)	R4 acc (%)	Total acc (%)	Time (s)
BPNN	0.42	85.38	78.00	76.25	30.00	81.64	2.15
SVM	0.52	86.00	72.67	73.75	50.00	80.78	4.25
RVM	0.57	79.88	79.67	75.00	65.00	78.98	1.95
Fast RVM	0.63	89.88	75.67	70.63	45.00	83.44	0.92

Compared with the BPNN, the SVM classifier has reduced dependence on training samples and obtains better G-mean values. The accuracy of class 4 is higher than that of the BPNN. However, the optimal hyperboundary surface in the SVM classifier favors the majority class (i.e. class 1 accuracy R1 acc = 92.63%) and results in the huge difference between major and minority classes. The better performance of the SVM over the BPNN indicates that the Structural Risk Minimization Principle for the SVM training can reduce the dependence of classifiers on the balance of data.

However, the application of RVM results in highest accuracy on class 4 compared with other classifiers, which matches the character of sparse probability models in the Bayesian learning framework, and the G-mean value is better than that for other classifiers.

Unfortunately, the performance of Fast RVM is comparably poor with the G-mean at 0.29 and class 4 accuracy at 22.22%, although the training time has been substantially reduced by approximately 50% compared with other classifiers.

The simulation results in Table 5 suggest that, for the four classifiers, the classification accuracies in class 4 are worse than the classification accuracies in class 1 because of the imbalance between the majority classes and the minority classes, indicating an unsatisfactory classification of the raw data. Among all of the classifiers, RVM performs best, but the G-mean remains quite low.

Classification performance on pre-processed data

The classification performances of classifiers on the pre-processed data are presented in Table 6.

Regarding the BPNN, the class 3 and class 4 accuracies have been raised from 60 to 76.25% and 5 to 30% compared with Table 5; while class 1 accuracy has been decreased from 92.88 to 85.38% and class 3 accuracy remains almost the same. However, the G-mean value has improved remarkably from 0.07 to 0.42 because of the obvious increased classification accuracies in expanded minority classes (class 3 and class 4).

Regarding the SVM classifier, the accuracies of class 2, class 3 and class 4 have been increased from 67.33, 61.25 and 30% to 76.67, 73.75 and 50%, respectively, with class 1 accuracy being declined from 92.63 to 83%, and the G-mean value has been raised from 0.39 to 0.52.

Regarding the RVM, the accuracies of class 2, class 3 and class 4 are all improved, but there is an obvious decrease of class 1 accuracy (from 91.88 to 79.88%) because of the dramatic decrease in data number after pre-processing. The G-mean of RVM has also been improved to 0.57.

Lastly, regarding the Fast RVM, class 4 accuracy has increased from 22.22 to 45%. Notably, class 1 accuracy is the best among all of the classifiers because the data reduction with Fast RVM has considered the fitness of classifying data onto Fast RVM classifiers. The G-mean value has been improved to 0.63 as the best among all of the classifiers; while the processing time is decreased to less than 1 second at 0.92 s.

We found the G-mean values of all of the classifiers improved notably as shown in Table 7, which highlights the value of data pre-processing. In the present case study, the multiple binary classifiers (i.e. SVM, RVM and Fast RVM) perform better than the multiple classifier BPNN and further study is needed for more in-depth discussions.

Table 7 | G-mean of the four classifiers with original data and processed data

G-mean	BPNN	SVM	RVM	Fast RVM
Original data	0.07	0.40	0.42	0.29
Processed data	0.42	0.52	0.57	0.63

CONCLUSIONS

This study demonstrates the application of a pre-processing method for imbalanced data classification with Fast RVM for under-sampling of majority class data and the SMOTE algorithm for over-sampling of minority class data. The method is applied to fault diagnosis in a WWTP. The case study suggests that this pre-processing method can neutralize the data imbalance between the majority and minority classes to increase the share of the minority classes among the whole database. For this case study, improvements for the four chosen classifiers (i.e. BPNN, SVM, RVM and Fast RVM) in G-mean values, classification times, and accuracy can be observed. The Fast RVM multi-classifier performs better compared with other classifiers in this certain case with time at 0.92 s and G-mean value at 0.63.

However, the present case study is limited, and it does not assert a general claim. Additionally, further research (including more case studies, data splitting, as well as models optimization) is required for more general conclusions. Finally, other data-preprocessing methods for imbalanced data should also be compared to develop recommendations for industrial application.

ACKNOWLEDGEMENTS

This work was supported by Science and Technology Planning Project of Guangdong Province, China (2016A020221008, 2016B090918028, 2016B090927007) and Science and Technology Planning Project of Guangzhou, China (201604010032). The authors are grateful to the anonymous reviewers and the editor for their helpful comments and suggestions which greatly helped us to improve the quality of the paper.

REFERENCES

Blagus, R. & Lusa, L. 2013 SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatic* **14** (1), 1–16.

Burges, C. J. C. 1996 Simplified support vector decision rules. In: *Proc. 13th International Conference on Machine Learning*, Bari, Italy, July 3–6, pp. 71–77.

Burges, C. J. C. & Scholkopf, B. 1997 Improving the accuracy and speed of support vector machines. In: *Advances in Neural Information Processing Systems 9*, Cambridge, MA, USA, December 2–5, pp. 375–381.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. 2002 SMOTE: synthetic minority over-sampling technique. *J. Art. Intell. Res.* **16** (1), 321–357.

Chen, M. C., Chen, L. S., Hsu, C. C. & Zeng, W. R. 2008 An information granulation based data mining approach for classifying imbalanced data. *Inf. Sci.* **178** (16), 3214–3227.

Cover, T. & Hart, P. 1967 Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13** (1), 21–27.

Fan, X. W. 2003 *Support Vector Machine and its Application*. PhD Thesis, Zhejiang University, China (in Chinese).

Fuente, M. J. D. L. & Vega, P. 1995 A neural networks based approach for fault detection and diagnosis: application to a real process. In: *Proc. 4th IEEE Conference on Control Applications*, Albany, NY, USA, September 28–29, pp. 188–193.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H. & Herrera, F. 2011 An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognit.* **44** (8), 1761–1776.

Gao, M., Hong, X., Chen, S. & Harris, C. J. 2011 A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems. *Neurocomputing* **74** (17), 3456–3466.

Hamed, M. M., Khalafallah, M. G. & Hassanien, E. A. 2004 Prediction of wastewater treatment plant performance using artificial neural networks. *Environ. Model. Soft.* **19** (10), 919–928.

Hartigan, J. A. & Wong, M. A. 1979 A K-means clustering algorithm. *Appl. Stat.* **28** (1), 100–108.

He, H. & Garcia, E. A. 2009 Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21** (9), 1263–1284.

Hong, Y. S., Rosen, M. R. & Bhamidimarri, R. 2003 Analysis of a municipal wastewater treatment plant using a neural network-based pattern analysis. *Water Res.* **37** (7), 1608–1618.

Hsu, C. W. & Lin, C. J. 2002 A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* **13** (2), 415–425.

Hu, J. G. & Hu, X. M. 2002 Management and automatically control in water treatment. *J. Wuhan Uni. Technol.* **24** (11), 66–67 (in Chinese).

Hwang, J. P., Park, S. & Kim, E. 2011 A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function. *Expert Syst. Appl.* **38** (7), 8580–8585.

Lee, D. S., Vanrolleghem, P. A. & Park, J. M. 2005 Parallel hybrid modeling methods for a full-scale cokes wastewater treatment plant. *J. Biotechnol.* **115** (3), 317–328.

Liu, J. S. 2011 *Research on Fast Classification Algorithm of Relevance Vector Machine Based on Clustering*. Master Thesis, South China University of Technology, China (in Chinese).

Liu, C. Z. & Han, J. Y. 2013 Application of support vector machine based on neighborhood rough set to sewage treatment fault diagnoses. *J. Gansu Agric. Uni.* **48** (3), 176–180 (in Chinese).

MathWorks 2013 MATLAB, 2013.3. Available from: https://www.mathworks.com/.

Motamarri, S. & Boccelli, D. L. 2012 Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms. *Water Res.* **46** (14), 4508–4520.

Nguyen, H. W., Cooper, E. W. & Kamei, K. 2011 Borderline over-sampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Par.* **3** (1), 4–21.

Polat, K. & Güneş, S. 2008 A novel data reduction method: distance based data reduction and its application to classification of epileptiform EEG signals. *Appl. Math. Comput.* **200** (1), 10–27.

Qian, Y. 2014 *Research on Application of Classification Algorithms for Imbalanced Data*. PhD Thesis, Jilin University, China (in Chinese).

Qian, Y., Liang, Y. C. & Guan, R. C. 2014 Improving activated sludge classification based on imbalanced data. *J. Hydroinform.* **16** (6), 1331–1342.

Raskutti, B. & Kowalczyk, A. 2004 Extreme re-balancing for SVMs: a case study. *ACM SIGKDD Exp. Newsl.* **6** (1), 60–69.

Rumelhart, D. E., Hintont, G. E. & Williams, R. J. 1986 Learning representations by back-propagating errors. *Nature* **323** (9), 533–536.

Sànchez, M., Cortés, U., Béjar, J., Grácia, J. D., Lafuente, J. & Poch, M. 1997 Concept formation in WWTP by means of classification techniques: a compared study. *Appl. Intell.* **7** (2), 147–165.

Tao, E. P., Shen, W. H., Liu, T. L. & Chen, X. Q. 2013 Fault diagnosis based on PCA for sensors of laboratorial wastewater treatment process. *Chem. Intell. Lab. Syst.* **128**, 49–55.

Tipping, M. E. 2001 Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **1** (3), 211–244.

Tipping, M. E. & Faul, A. 2003 Fast marginal likelihood maximization for sparse Bayesian models. In: *Proc. 9th International Workshop on Artificial Intelligence and Statistics*, Key West, FL, USA, January 3–6, pp. 3–6.

University of California Irvine (UCI) 1987 *University of California Irvine Machine Learning Repository*. Available from: http://archive.ics.uci.edu/ml/datasets/Water+Treatment+Plant.

Vapnik, V. N. 1995 *The Nature of Statistical Learning Theory*. Springer, New York.

Yang, Z. R. 2006 A fast algorithm for relevance vector machine. In: *Proc. 7th International Conference on Intelligent Data Engineering and Automated Learning*, Burgos, Spain, September 20–23, pp. 33–39.

Yen, S. J. & Lee, Y. S. 2009 Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Syst. Appl.* **36** (3), 5718–5727.

Zhang, Y. L. & Zhu, Y. W. 2011 Recognition of waste water class based on support vector machine. *J. Tangshan Teach. Coll.* **33** (2), 30–32 (in Chinese).

Zong, W. W., Huang, G. B. & Chen, Y. Q. 2013 Weighted extreme learning machine for imbalance learning. *Neurocomputing* **101** (3), 229–242.

First received 15 October 2015; accepted in revised form 1 November 2016. Available online 6 January 2017