Improving activated sludge classification based on imbalanced data

Y. Qian, Y. C. Liang and R. C. Guan

ABSTRACT

A fast and accurate classification method for sewage sludge biological activity classification is of great significance for wastewater treatment. However, the data are often imbalanced and the accuracy of traditional classification algorithms applied to imbalanced small classes of data is very low. Such small classes are crucial application data. Therefore, based on the analysis of eight microorganisms, a novel method is proposed in this paper for the classification of activated sludge known as balanced support-vector-based back-propagation (SV-BP) neural network. It first splits the multiclass classification problem into a plurality of pairwise classification problems and uses a support vector machine (SVM) to achieve equalization. Second, the new dataset is produced, following which back-propagation neural network (BPNN) is used for training and classification. To examine the efficiency of the model, 1731 real data points are collected from a wastewater treatment factory and divide the data into four classes with the help of wastewater experts. Based on the new model, data redundancy and noise are greatly reduced. With area under the curve (AUC) measurements, we find that the AUC of SV-BP is 6.9% higher than classical BPNN. In addition, the small-class recognition rate of SV-BP is far better than that by classical BPNN and SVM algorithms.

Key words | activated sludge classification, back-propagation neural network, imbalanced data, support vector machine

Y. Qian
Y. C. Liang
R. C. Guan (corresponding author)
Key Laboratory for Symbol Computation and Knowledge Engineering of National Education Ministry,
College of Computer Science and Technology, Jilin University,
Changchun 130012,
China
E-mail: *guanrenchu@jlu.edu.cn*

Y. Qian
College of Electrical and Information Engineering,
Beihua University,
Jilin 132021,
China

R. C. Guan
State Key Laboratory of Inorganic Synthesis and Preparative Chemistry,
Jilin University,
Changchun 130012,
China

INTRODUCTION

Motivation

Activated sludge processing is one of the most common wastewater treatments. In this paper, it mainly contains four parts: primary settler, hydrolysis acidification tank, biological tank and secondary settler. The raw wastewater composed of industrial, neutralized, and domestic wastewater firstly flows into primary settler and then is transported to the hydrolysis acidification tank. From the flowchart shown in Figure 1, it can be seen that the key principle underlying this kind of sewage treatment is the third part which is named as biological tank. It is to use the metabolism of microorganisms to degrade wastewater pollutants, such as organic matter and disease-causing microorganisms. To improve the efficiency of this method, a direct approach is to use high-quality activated sludge, which is measured by its biological activity. By analysing the bioactivity of microorganisms in the biological tank, wastewater treatment experts can judge the quality level of the sludge bioactivity and sewage treatment. For hundreds of wastewater plants and data, however, human experts may misclassify the activated sludge. This also requires expensive human labour and time.

The state-of-the-art

Therefore it is very important to build a good mathematical model that could satisfy this requirement of wastewater treatment plants (WWTP). One of the most popular conceptual models is the Activated Sludge Model No. 1 (ASM1) (Henze *et al.* 1987), and its more complex versions ASM2d (Henze *et al.* 1999) and ASM3 (Gujer *et al.* 1999). However,

doi: 10.2166/hydro.2014.123

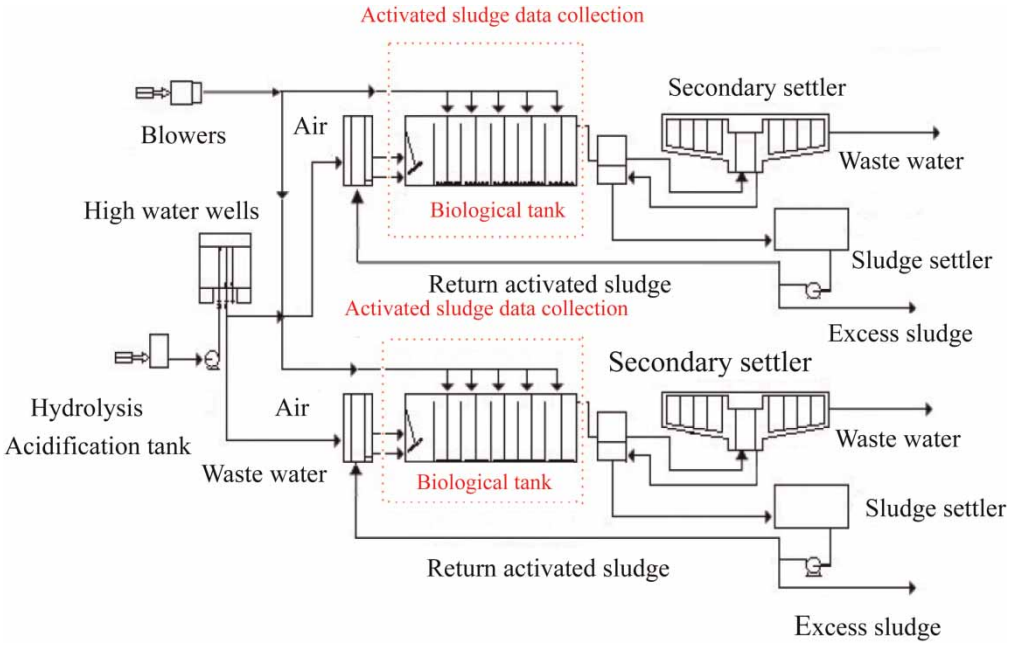

Figure 1 | Flowchart of sewage treatment process.

the components of wastewater are numerous and varied, and the related domain specific knowledge is very complex. Therefore, it can result in increasing complexity of the mathematical models.

Recently, with the fast development of machine learning, some effective methods have been used to solve practical problems. For example, Atanasova & Kompare (2002) proposed a decision trees model to predict WWTP operation; Chawla & Hunter (2005) proposed that bathing water quality can be classified via the Hazen method; Kompare *et al.* (2006) implemented the well-known conceptual model ASM1 to simulate pilot WWTP processes and a regression trees model to predict ammonia outflow concentration; Shamseldin developed a real-time river flow forecasting model using artificial neural networks (Shamseldin 2010); Yang *et al.* discussed three feature selection techniques (information gain, mutual information and relief), and tested them on a sustainable flood retention basins dataset based on support vector machines (SVMs), k-nearest neighbours, C4.5 decision trees and naive Bayes classification techniques (Yang *et al.* 2011); Mannina *et al.* (2011) proposed a procedure for the calibration of an activated sludge model based on comprehensive sensitivity analysis and a novel step-wise Monte Carlo-based calibration of the subset of influential parameters; Motamarri & Boccelli (2012) classified the level of recreational water quality with fecal indicator organisms, using multivariate linear regression and artificial neural networks.

Imbalanced data problem

However, our problem, i.e. the quality of activated sludge, can often be classified into four categories: excellent, good, general and poor. The different categories have greatly imbalanced distributions, for example, samples in 'general' and 'good' classes are generally much larger than those in 'poor' and 'excellent' categories (in our case, the former two categories are approximately 40 times more prevalent than that of the latter two). This imbalance introduces great difficulties in the automatic classification of wastewater sludge and prediction of biological activity. For multiple discriminant analysis, it is generally believed that when the proportion of the minority and the majority class falls below 1:2, the data are imbalanced (Li 2011). This is one of the main challenges in data mining and machine learning research. Imbalanced data exist in many real applications (He & Garcia 2009), such as biomedical research (Eitrich *et al.* 2007; Li *et al.* 2010), credit card fraud detection

(Chan *et al*. 1999), market analysis of business conduct (Bose & Chen 2009; Xiao *et al*. 2012), and Web mining.

The difficulty of imbalanced data processing is that most conventional classification algorithms assume that the prior probability distribution of samples is a discrete uniform distribution or that the misclassification cost is equivalent. However, when faced with non-uniformly distributed data, the small category samples will be drowned by larger category samples. Error rates from small category classification are relatively higher than those from large ones.

Since the late 1960s, researchers began to study the imbalanced data classification problem. Such research can be grouped into two distinct points of view. First, from the viewpoint of the data, under-sampling and over-sampling techniques are proposed to achieve a balanced distribution of data. For instance, Cover & Hart (1967), developed an under-sampling algorithm based on a condensed nearest neighbour (CNN) training sample reduction approach, but this algorithm is sensitive to its initial values, and class boundaries are foggy. This is attributed to redundant training samples, as well as boundary training samples misleading the decision. An extended CNN-based algorithm was therefore proposed by Hao & Jiang (2007). Its training data were filtered by voting, which effectively reduces redundant data in the training set but leaves behind boundary ambiguities. Wilson (1972) proposed another under-sampling method by editing nearest neighbour results, which can effectively reduce the influence of wrong classification boundary samples. Conversely, Chawla *et al*. (2002) proposed an over-sampling algorithm to construct small-class samples.

Second, from the viewpoint of the algorithms, Chen *et al*. (2008) established knowledge acquisition on imbalanced datasets using back-propagation neural networks (BPNNs). Oh (2011) proposed another back-propagation algorithm based on the error function to adjust the weights of large and small classes. To improve accuracy and reduce training time, Zhao (2009) proposed an artificial neuron algorithm to perform imbalanced data classification. López *et al*. (2012) drew the conclusion that pre-processing and cost-sensitive learning can both address the imbalance problem quite well; however, to the best of our knowledge, there is no relevant study on the typical imbalance data classification problem of sludge biological activity.

In this paper, based on the data analysis of a sewage treatment plant in China, a novel hybrid algorithm involving a balanced support-vector-based back-propagation (SV-BP) neural network is proposed. The new approach uses SVM to find support vectors (SVs) in the raw and multi-categories data to construct a new training set. Then, the BP network is trained on this new training set to obtain the final classifier. Both traditional BP neural networks and SVMs were applied to the same dataset to compare some traditional methods with our new algorithm. It can be found that the SV-BP classification algorithm performed much better than traditional ones on both classification accuracy and minority category recognition rates.

BACKGROUND

Back-propagation neural network

BPNNs were first proposed by Rumelhart *et al*. (1986). BPNNs are not only good at parallel and distributed processing, nonlinear mapping, generalization and fault tolerance, but they are also able to attain self-learning, self-organizing and self-adaptive capacities. Until now, BPNNs have been widely applied to many fields, such as environmental engineering (Wen & Vassiliadis 1998) and industrial automation (Zhang & Stanley 1999). As shown in Figure 2, the network includes an input layer, a hidden layer and an output layer. The basic idea of BPNN is that the learning process consists of both forward propagation and backward propagation of errors. If forward propagation output does not satisfy the predefined expectation, the errors would propagate back to adjust the weights. When the algorithm converges, it would identify the best weights, making the BPNN achieve the least amount of error.

Support vector machines

SVMs were first proposed by Cortes & Vapnik (1995). The basic idea of an SVM is to use structural risk minimization instead of traditional empirical risk minimization. SVMs are based on a well-founded theoretical approach and can converge to the global and unique optimum and have

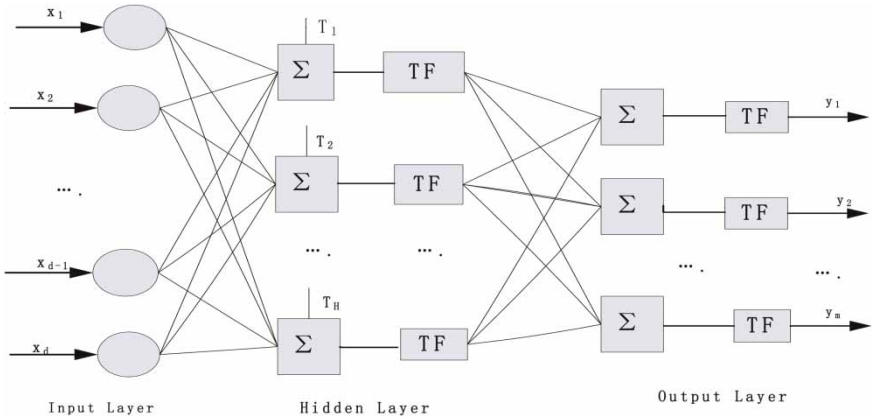

Figure 2 | Topology of BPNN.

hence been successfully applied in many fields, including pattern recognition (Liu & Chen 2007), fault diagnoses (Widodo & Yang 2007) and bioinformatics (Zhang *et al.* 2009). Its mathematical model can be depicted as described below.

For binary classification, given training set $\{(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)\}$, $x_i \in R^d$, where n is the number of training samples, x_i is a d-dimensional vector and $y_i = \{-1, +1\}$ is its category label. SVM finds classification hyperplane H with the most powerful generalization ability and maximum distance between two categories as $H{:}(w \cdot x) + b = 0$, where w and b are the weight vector and threshold of the function, respectively. Then, the nearest sample point x_i to hyperplane H should satisfy the conditions (interval $\delta = 1$)

$$\left.\begin{array}{l} H_1{:}(w \cdot x_i) + b = 1,\ y_i = 1 \\ H_2{:}(w \cdot x_i) + b = 1,\ y_i = -1 \end{array}\right\}. \tag{1}$$

Assuming the distance between point and classification interface is d, then

$$d = \frac{|(w^T x) + b|}{\|w\|} = \frac{1}{\|w\|}. \tag{2}$$

Therefore, the sum of the interval between the two categories is the margin of $2/\|w\|$, which is shown in Figure 3.

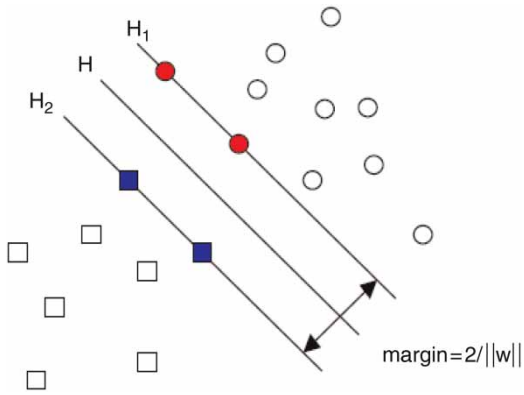

Figure 3 | Optimal classification interface.

Using the dual problem $\min\|w\|^2$ instead of $\max(2/||w||)$, equation (3) is generated as

$$\left.\begin{array}{l} \min F(w, b) = \min \dfrac{1}{2}\|w\|^2 \\ \text{st: } y_i(w^T x_i + b) - 1 \geq 0,\ i = 1, 2, \ldots n \end{array}\right\}. \tag{3}$$

To find the conditional extreme value, the extended Lagrange function is introduced as

$$\left.\begin{array}{l} L(w, b, \alpha) = \dfrac{1}{2}\|w\|^2 - \displaystyle\sum_{i=1}^{n} \alpha_i y_i (w^T \cdot x_i + b) + \sum_{i=1}^{n} \alpha_i \\ \alpha_i \geq 0, \quad i = 1, 2 \ldots n \end{array}\right\}, \tag{4}$$

where α_i is the Lagrange multiplier, $L(w, b, \alpha)$ is derived from partial evaluation and w and α are limited

as follows:

$$\left.\begin{array}{l} w-\sum\limits_{i}\alpha_i y_i x_i=0 \\ \sum\limits_{i=1}^{n}\alpha_i y_i=0 \end{array}\right\}. \tag{5}$$

Updating Equations (4) with (5), the dual problem of formula (3) turns out to be

$$\left.\begin{array}{l} \max\ W(\alpha)=\sum\limits_{i}\alpha_i-\dfrac{1}{2}\sum\limits_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i x_j \\ \sum\limits_{i=1}^{n}\alpha_i y_i=0;\ \alpha_i\geq 0;\quad i=1,2\ldots n \end{array}\right\}. \tag{6}$$

The solution to Equation (6) is the optimal solution of original optimization Equation (3). The vectors corresponding to $\alpha_i > 0$ are the SVs; they are the data close to the hyperplane.

Data preprocessing method for imbalanced data

Farquad & Bose (2012) successfully applied SVM to solve a binary imbalanced data problem. They first extracted SVs from the data and used the new dataset instead of the original data. Their experiment on insurance data shows a larger improvement on sensitivity. The two imbalanced classes consist of a distribution of 94:6. With 25% sampling and 50% preprocessing, a comparatively balanced dataset was produced. Finally, traditional multilayer perceptron, random forest and logistic regression algorithms were applied to compare the accuracy before and after pre-processing. In this paper, this methodology is extended to address the specific domain, namely sewage sludge biological activity classification; further, the original idea is extended from two categories to multicategories of data and a hybrid approach that integrates preprocessing and cost-sensitive learning is proposed.

BPNN BASED ON BALANCED SUPPORT VECTORS

Preprocessing

Classification using all of the imbalanced samples will not only increase the amount of required training time, but also cause a decrease in the generalization ability in the 'over-fitting' phenomenon and therefore decrease classification accuracy. To improve the balance ratio of multi-categorisation, we employ the SVM algorithm is employed, which includes two steps described below.

First, the original training set is assumed to consist of M categories, which are $\omega_1, \omega_2 \ldots \omega_M$, and the sample numbers are $S_1, S_2, \ldots, S_M$. Next, every pair of classes ω_i and ω_j is taken to construct a classifier; therefore it has a total of $(M-1)\times M/2$ classifiers. The corresponding number of discriminant function bias b is $(M-1)\times M/2$. For each sample, its classification result depends on the competition of all classifiers. Thus, the classifier optimization problem is transformed from Equation (3) into

$$\left.\begin{array}{l} \min\ F\left(w_{ij}, b_{ij}, \xi_{ij}\right)=\dfrac{1}{2}\|w_{ij}\|^2+C\sum\limits_{t=1}^{n}\xi_{ij} \\ \text{st: } \xi_{ij}\geq 0 \\ \text{st: } \left(\left(w_{ij}\right)^T\phi(x_t)\right)+b_{ij}\geq 1-\xi_{ij},\ \text{if: } x_t\in\omega_j \\ \text{st: } \left(\left(w_{ij}\right)^T\phi(x_t)\right)+b_{ij}\leq -1+\xi_{ij},\ \text{if: } x_t\in\omega_j \end{array}\right\}. \tag{7}$$

Second, if $S_i \gg S_j$, it is an imbalanced distribution problem. Suppose ω_i is the majority class (the negative class) and ω_j is the minority class (the positive class); to eliminate the impact of the imbalanced distribution, the penalty factors is introduced to adjust the positive and negative classes C_j^+ and C_i^- for each classifier. With the introduced penalty, the SVM training procedure becomes

$$\left.\begin{array}{l} \min\ F\left(w_{ij}, b_{ij}, \xi_{ij}\right)=\dfrac{1}{2}\|w_{ij}\|^2+C_j^+\sum\limits_{y_j=1}\xi_{ij}+C_i^-\sum\limits_{y_i=-1}\xi_{ij} \\ \text{st: } \xi_{ij}\geq 0 \\ \text{st: } \left(\left(w_{ij}\right)^T\phi(x_t)\right)+b_{ij}\geq 1-\xi_{ij},\ \text{if: } x_t\in\omega_j \\ \text{st: } \left(\left(w_{ij}\right)^T\phi(x_t)\right)+b_{ij}\leq -1+\xi_{ij},\ \text{if: } x_t\in\omega_j \end{array}\right\}, \tag{8}$$

where C_j^+ is the penalty of misclassifying a negative class sample as a positive sample and C_i^- is the penalty of misjudging a positive class sample as a negative sample.

Then, the dual form becomes

$$\left.\begin{array}{l}\min\limits_{\alpha} \dfrac{1}{2}\alpha^T Q\alpha - \mathrm{e}^T\alpha \\ \text{st: } 0 \le \alpha_j \le C_j^+,\ y_j = 1 \\ \text{st: } 0 \le \alpha_i \le C_i^+,\ y_i = -1 \\ \text{st: } y_t\ \alpha = 0\end{array}\right\}, \tag{9}$$

where e is the unit vector and Q is a positive semi-definite matrix. Based on Chang & Lin (2011), further consolidation yields

$$\left.\begin{array}{l}\min\limits_{\alpha_i,\alpha_j} \dfrac{1}{2}[\alpha_i,\alpha_j]\begin{bmatrix}Q_{ii}, & Q_{ij}\\ Q_{ji}, & Q_{jj}\end{bmatrix}\begin{bmatrix}\alpha_i\\ \alpha_j\end{bmatrix} + (Q_{i,N}\alpha_N - 1)\alpha_i + (Q_{j,N}\alpha_N - 1)\alpha_j \\ \text{st: } y_i\alpha_i + y_j\alpha_j = \Delta - y_N^T\alpha_N^k \\ \text{st: } 0 \le \alpha_i \le C_i^- \\ \text{st: } 0 \le \alpha_j \le C_j^+\end{array}\right\}. \tag{10}$$

Let $\alpha_i = \alpha_i^k + d_i$, $\alpha_j = \alpha_j^k + d_j$, $\hat{d}_i \equiv y_i d_i$ and $\hat{d}_j \equiv y_j d_j$; then Equation (10) becomes

$$\left.\begin{array}{l}\min\limits_{d_i,d_j} \dfrac{1}{2}[d_i,d_j]\begin{bmatrix}Q_{ii}, & Q_{ij}\\ Q_{ji}, & Q_{jj}\end{bmatrix}\begin{bmatrix}d_i\\ d_j\end{bmatrix} + \left[\nabla f\left(\alpha^k\right)_i, \nabla f\left(\alpha^k\right)_j\begin{bmatrix}d_i\\ d_j\end{bmatrix}\right] \\ \text{st: } y_i d_i + y_j d_j = 0 \\ \text{st: } -\alpha_i^k \le d_i \le C_i^- - \alpha_i^k \\ \text{st: } -\alpha_j^k \le d_j \le C_j^- - \alpha_j^k\end{array}\right\}. \tag{11}$$

Equation (11) is the optimal solution of original problem (8). Its decision function is shown as in Equation (6), where the corresponding vectors of $\alpha_i > 0$ are the SVs.

The classification model

To effectively solve our imbalanced multi-category activated sludge classification problem, we propose a balanced SV-BP neural network. The area under the curve (AUC) is employed as our evaluation standard to measure the results. The detailed process is shown in Figure 4.

Evaluation

Most traditional classification algorithms assume category distribution is balanced; therefore the performance evaluation method uses the overall recognition rate. However,

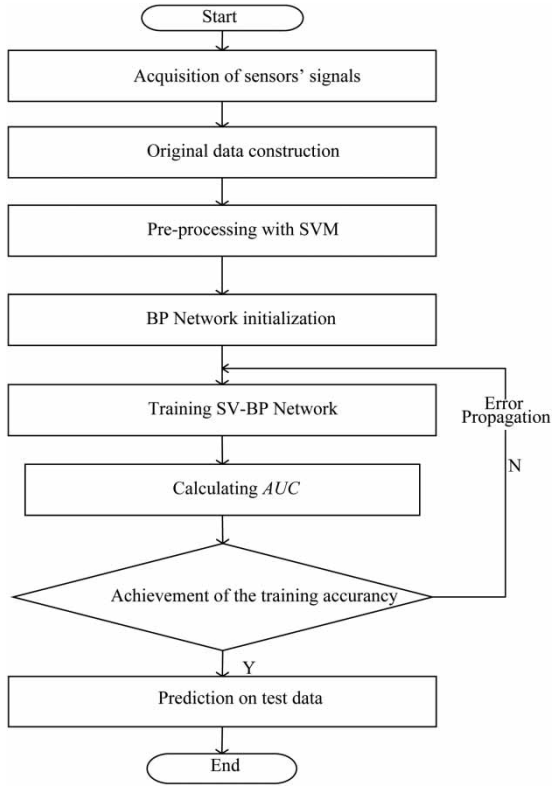

Figure 4 | Flowchart of proposed classifier.

for imbalanced datasets, classical evaluation methods often lead to high recognition rates of the majority class, whereas the recognition rates for the minority class are low. This result cannot adequately satisfy the demand for high accuracy on the minority class (e.g. the 'excellent' sludge recognition rate) in actual production. Therefore, traditional evaluation methods are not suitable for imbalanced data. There are other evaluation methods to address this problem, including for example, F-measure, G-mean, receiver operating characteristic curve (ROC), AUC and optimized precision (Ranawana & Palade 2006). For activated sludge classification, the weighted AUC-w is employed as evaluation method. The mathematical description is as follows. Let the negative sample set be $S^- = \{S_1^-, S_2^-, \ldots S_{M1}^-\}$, where $M1$ is its category number, and let the positive sample set be $S^+ = \{S_1^+, S_2^+, \ldots S_{M2}^+\}$, where $M2$ is the category number. Then, negative sample set element S_i^- and positive sample set element S_i^+ are randomly selected to construct a mixed matrix of binary classification. To pay more attention to the accuracy of the

minority class, the higher weight is given. Then, AUC is used to measure the classification results and the model is

$$\max F(W) = \sum_{i=1}^{M1} \sum_{j=1}^{M2} W_{ij} AUC_{ij}, \quad (12)$$

where i indicates the negative class and j represents the positive class. To look for maximum $F(W)$, the training parameters of the classification algorithms are optimized to achieve the best performance. To evaluate classifier learning performance, sensitivity, the overall recognition rate and specificity are also employed as evaluations measures for our proposed method.

EXPERIMENTS AND ANALYSIS

Dataset

The dataset is obtained from an activated sludge data collection system for a large biological wastewater treatment plant (Figure 1). The ecosystem in the activated sludge includes bacteria, protozoa and metazoans. There is a complicated competition for survival and ecological balance among these microbial populations. They form a symbiotic relationship with great adaptation to the extreme environments. When environmental parameters such as water quality and treatment procedure are fixed, the activated sludge will develop into a stable biological community. For example, if a large amount of toxic substance flows into sewage sludge, the strong-tolerance species will grow rapidly, and their population will be much larger than that of other species.

According to rules provided by activated sludge experts, indicators involving eight groups can fully describe sludge quality. These eight groups are Mastigophora (X_1), Amoebae (X_2), *Vorticella* (X_3), *Epistylis* (X_4), *Opercularia* (X_5), Suctorida (X_6), rotifers (X_7) and *Trachelophyllum* (X_8). From the expert rules shown in Table 1, it can be noted that these microbial populations can reflect the activated sludge activity, but they lack accurate quantitative relationships. Therefore, these eight microbial populations' data are collected from the aeration tank of a chemical plant in China from January 2007 to December 2008. Excluding maintenance time, 1731 data points which cover two years are obtained. The monthly average density of eight kinds of microorganisms is shown in Figure 5, of which the density unit is ind./ml, representing the total numbers of such micro-organisms per milliliter of sludge.

To evaluate the state of biological activity in activated sludge, the classical k-means clustering technique and wastewater expert knowledge are employed. The k-means algorithm is good at natural cluster detection. During the clustering process, different cluster scales were run to pursue the best clustering results. The number of clusters C and the clustering error squares are shown in Figure 6, where point A represents the dramatic drop in the error squares and the time complexity ($O{:}k \times n \times t$, where n is the sample number and t is the number of iterations) is

Table 1 | Classification rules of human experts

No.	If (condition)	Then (state of activated sludge)
1	Mastigophora (X_1) and other protozoa appeared in large numbers, and Peritrichida are rare	Bad
2	Amoebae (X_2) and rotifers (X_7) appeared in large numbers	Bad (dispersion or dissolution)
3	There are large number of Mastigophora (X_1) and the number of bare amoebae (X_2) is small	Bad (high load or the existence of refractory material)
4	*Vorticella* (X_3), *Epistylis* (X_4), *Opercularia* (X_5), Suctorida (X_6) and other sessile ciliated classes and creeping type animals appeared in large numbers and in good shape and also account for the entire biological individual number 80% or more	Excellent (water clarification)
5	*Trachelophyllum* (X_8) and other slow swimming-type protozoa appeared in large numbers	Bad–> good
6	Highly diverse microbial species, there is no dominant number of microorganisms	Good
7	Few microbial species, or a microorganism dominant	Bad

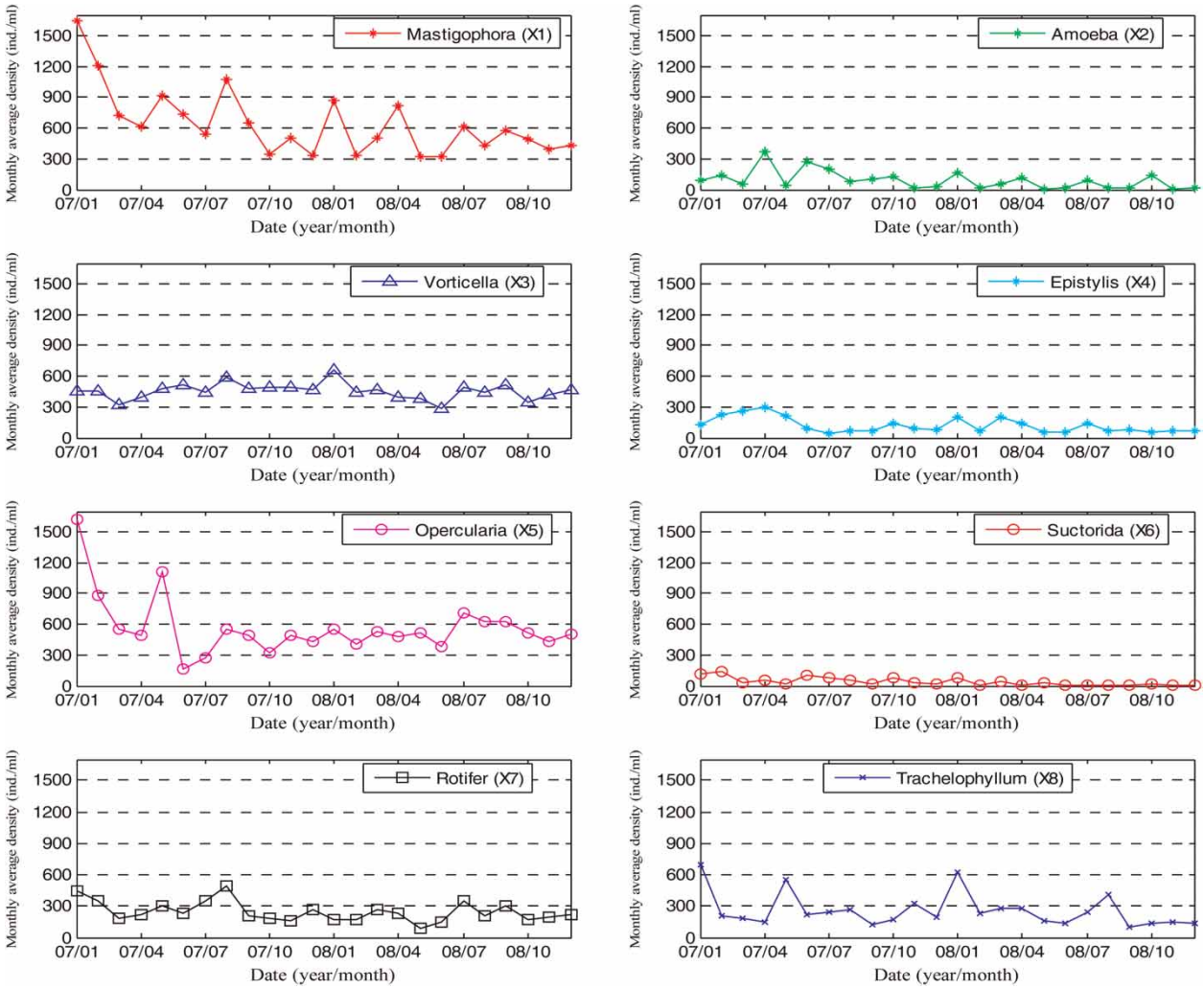

Figure 5 | The monthly average density of eight kinds of microorganisms.

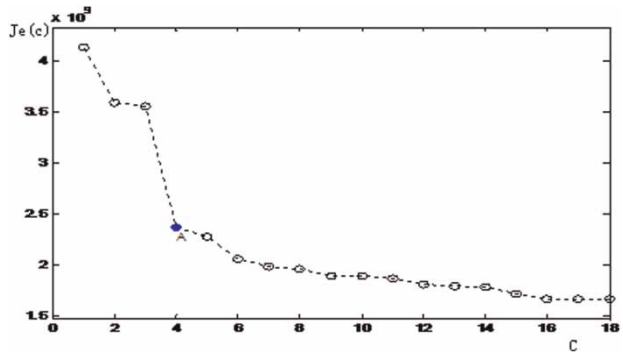

Figure 6 | Cluster C and the cluster's error square.

smaller than the points behind A on the tail. Based on these data, the cluster count and sludge bioactivity quality level are set as four. Fortunately, with the knowledge of sludge bioactivity quality analysis experts in the plant, the category number coincides with the following four grades: excellent, good, general and poor.

The more serious problem is that the class distribution of each category is 20:292:1393:26, which is shown in the first line of Table 2. It can be noted that this dataset presents a typical instance of the imbalanced distribution problem. In the experimentation, the 'excellent', 'good' and

Table 2 | Distribution of datasets

	Excellent	Good	General	Poor	Total
Original dataset	20	292	**1,393**	26	1,731
Original training set	12	172	**662**	19	865
Validation set	8	120	**731**	7	866
New training set	**12**	**125**	**191**	**18**	**346**

'poor' classes correspond to the minority class (positive) sample; the 'general' grade corresponds to the majority class (negative).

Of the 1,731 data points, 865 data points (collected in 2007) are considered as the training set, while the remaining 866 data points (collected in 2008) are served as the validation set. Detailed information regarding the dataset is shown in Table 2. From the third line of the table, it can be observed that the ratios of category 1 (excellent), category 2 (good) and category 4 (poor) account for only 1.81, 25.98 and 2.87% of category 3. In other words, category 3 (general) is approximately 54, 3 and 34 times larger than that of the aforementioned three classes. These data indeed show the imbalance of activated sludge data.

Parameter optimization

To avoid the negative influence outliers of a large numerical interval can have, the original data were normalized to the range [−1, +1]. Then, a reasonable nonlinear kernel is selected to extract the SVs and handle the complex relationships among the microbial populations in their specific ecosystem. Keerthi & Lin (2003) proved that a polynomial kernel function is a special state of a radial basis function (RBF) and Lin & Lin (2003) indicated that in some conditions, the sigmoid kernel and RBF kernel have similar features. Therefore, as it could not get enough prior knowledge, RBF is selected as kernel function $K(x_i, x_j) = \exp(-\gamma/\|(x_i - x_j)\|^2)$; however, the optimal values for penalty factors C and γ of SVM also need careful selection. Therefore, the LibSVM-grid-search tool developed in 2011 by Professor Lin of Taiwan University (Chang & Lin 2011) is utilized to search for the best values for the penalty factors. Its training set is shown as the third row of Table 2, while the search space is shown in Figure 7. It can be observed that when the parameter search process achieves the optimal

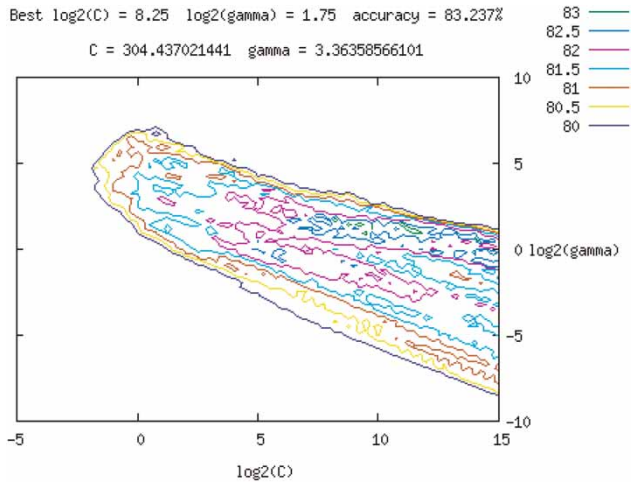

Figure 7 | Searching for the best parameter values of C and γ.

values, $C = 304.44$ and $\gamma = 3.36$, the classification accuracy rate of five-fold cross-validation is Acc = 83.237%.

When searching the SVs, because the number of categories is four, our multiclass classifier is first split into a plurality of two classifiers. Therefore, $p = (4-1) \times 4/2 = 6$ binary classifiers are constructed which are shown in Table 3. In Table 3, class 1, 2, 3 and 4 respectively indicate the classes of 'excellent', 'good', 'general' and 'poor' and the classification hyperplane parameters are shown in the second line. The number of SVs on the hyperplane interface is shown in the third line. From Table 2, it can be seen that 346 SVs were found to form the new training set, which was reduced from the original 865. Among these SVs, the 'excellent' and 'general' ratio was changed from the original 12:662 to 12:191; the 'poor' and 'general' ratio was changed from the original 19:662 to 19:191; and the 'good' and 'general' ratio was changed from the original 172:662 to 125:191. From Table 3, it can be observed that the count of samples in class 3 (general) and class 2 (good) significantly decreased to the ratio of 71.15 and 27.33%, respectively. The new distribution of the training set is 12:125:191:18, which is more balanced than the original data. Experimental results show that the noise and redundancy in the new training set were reduced and the classifier training time was also sharply decreased.

After preprocessing, BP network is selected to do the classification, here d is set as the number of input nodes, H as the number of hidden layer nodes and m as the number

Table 3 | Separating hyperplane parameters

Hyperplane	Class 3 and 2	Class 1 and 3	Class 3 and 4	Class 2 and 1	Class 2 and 4	Class 1 and 4
Bias	− 1.82	− 1.98	− **8.06**	− 1.70	− 0.96	− 0.77
Number of SVs	**252**	39	64	29	36	15
Original training set	662:172	**12:662**	662:19	172:12	172:19	12:19
New training set	191:125	**12:191**	191:19	125:12	125:18	12:18

Class 1: excellent; Class 2: good; Class 3: general; Class 4: poor.

Table 4 | Accuracy rate (%) and AUC results of SVM, BP and SV-BP

Algorithm	Total RR*	Class 1 AR	Class 2 AR	Class 3 AR	Class 4 AR	Parameters
SVM	83.72	0.00	0.83	99.04	0.00	AUC1 = 1.98; T1 = 350 s
BP	84.41	0.00	0.00	**100.00**	0.00	AUC2 = 1.97; T2 = 368 s AUC2 = 1.97
SV-BP	**84.53**	**50.00**	**15.00**	97.13	0.00	AUC3 = **2.06**; T3 = **220 s** AUC3 = 2.06

Total RR is the total recognition rate; Class X AR indicates the accuracy rate of class X(X = 1, 2, 3, 4); AUC1, AUC2 and AUC3 represent AUC of SVM, BP, and SV-BP algorithms; T1, T2 and T3 represent runtime of SVM, BP, and SV-BP algorithms, respectively.

of output nodes. The topology of the network is the 'N–H–M' three-tier structure, as shown in Figure 2 above. Eight micro-organism indicators were taken as input nodes; therefore, the feature number was $d = 8$. Again, the four categories of 'excellent', 'good' , 'general' and 'poor' were used as output nodes. The hyperbolic tangent S is used as the transfer function in the hidden and output layers. The range of H was set as 11 based on the experimental experience.

COMPARISON AND DISCUSSION

Classification results of the SVM, BP and SV-BP classifiers are summarised in Table 4 where 'Total RR' indicates total recognition rate and 'Class X AR' is the accuracy rate of the Xth class. Results show that SV-BP is better than traditional SVM and BP algorithms in both the total recognition accuracy rate and the small-class recognition rate. Especially for the latter measure, the first small class (i.e. 'excellent') even cannot be recognized in the BP and SVM approaches, but SV-BP achieves 0.5. The runtime is also reduced to one-third of that of traditional algorithms.

To overcome the misclassification costs for each category, AUC is used as a measurement. Class 3 (general) indicates the majority class, and all the others (class 1, class 2 and class 4) are the minority classes in this paper. The ROC simulation data are shown as the last column of Table 4. Larger values of AUC mean that the corresponding classifiers are more concerned about minority class accuracy, so the classifier performance in the third row (AUC3 = 2.06) is better than that of the second row (AUC2 = 1.97) of Table 4. Therefore, it can be observed that when using the traditional BP algorithm, it clearly assumes that the prior probability distribution of each class is balanced; however, the minority class samples error rate was higher. Conversely, SV-BP overcame the adverse effects experienced by the BP algorithm due to imbalanced data, because the extraction of SVs by SVM established a relatively balanced dataset. A classifier based on our SV-BP algorithm is not only effective in removing redundant information from the training set, but also increases the minority class recognition rate, increases prediction accuracy and reduces runtime.

CONCLUSIONS

In this paper, a novel approach for automatically classifying activated sludge is proposed. First, eight microbial groups were used to construct a feature space based on expert knowledge, then employed k-means to discover different

clusters of sludge quality. Further, to solve the most difficult problem, i.e. significantly multi-classes imbalanced data, a novel SV-BP algorithm is proposed. To evaluate our new algorithm's performance, traditional SVM and BP networks are compared with SV-BP. Simulation results showed that the SV-BP algorithm not only effectively removed information redundancy and noise but also obtained higher accuracy and reduced classifier training time. In addition, the recognition rates of both the overall performance and minority classes were better than that of traditional approaches. This new pipeline can therefore improve the ability to automatically classify quality levels of activated sludge. This in turn will help realize better monitoring of sludge processing conditions at biological treatment and improved controlling of the production requirements of return sludge and discharge volumes, achieving the overall purpose of energy conservation.

ACKNOWLEDGEMENTS

The authors are grateful for the support of the National Natural Science Foundation of China (grant nos 61103092, 41101376, 61073075, 61272207), the China Postdoctoral Science Foundation (grant nos 2011M500613, 2012T50298), the Science Technology Development Project from Jilin Province (grant nos 20120730, 20130522106JH, 20140520070JH) and the PhD Program Foundation of MOE of China (no. 20120061110094).

REFERENCES

Atanasova, N. & Kompare, B. 2002 The use of decision trees in the modelling of a wastewater treatment plant. *Acta Hydrotech.* **20** (33), 351–370.

Bose, I. & Chen, X. 2009 Hybrid models using unsupervised clustering for prediction of customer churn. *J. Organ. Comput. Electr. Commun.* **19** (2), 133–151.

Chan, P. K., Fan, W., Prodromidis, A. L. & Stolfo, S. J. 1999 Distributed data mining in credit card fraud detection. *IEEE Intell. Syst. Appl.* **14** (6), 67–74.

Chang, C. C. & Lin, C. J. 2011 LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2** (3), 1–27.

Chawla, R. & Hunter, P. R. 2005 Classification of bathing water quality based on the parametric calculation of percentiles is unsound. *Water Res.* **39** (18), 4552–4558.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. 2002 SMOTE: synthetic minority over-sampling technique. *J. Artif. 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.

Cortes, C. & Vapnik, V. 1995 Support-vector networks. *Mach. Learn.* **20** (3), 273–297.

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

Eitrich, T., Kless, A., Druska, C., Meyer, W. & Grotendorst, J. 2007 Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. *J. Chem. Inf. Model.* **47** (1), 92–103.

Farquad, M. A. H. & Bose, I. 2012 Preprocessing unbalanced data using support vector machine. *Decis. Support Syst.* **53** (1), 226–233.

Gujer, W., Henze, M., Mino, T. & Loosdrecht, M. V. 1999 Activated sludge model no. 3. *Water Sci. Technol.* **39** (1), 183–193.

Hao, H. W. & Jiang, R. R. 2007 Training sample selection method for neural networks based on nearest neighbor rule. *Acta Autom. Sin.* **33** (12), 1247–1251.

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

Henze, M., Grady Jr., C. P. L., Gujer, W. & Matsuo, T. 1987 Activated Sludge Model No. 1. Scientific and Technical Reports No. 1, IAWPRC, London.

Henze, M., Gujer, W., Mino, T., Matsuo, T., Wentzel, M. C. & Van Loosdrecht, M. 1999 Activated sludge model no. 2d, ASM2d. *Water Sci. Technol.* **39** (1), 165–182.

Keerthi, S. S. & Lin, C. J. 2003 Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Comput.* **15** (7), 1667–1689.

Kompare, B., Levstek, M. & Atanasova, N. 2006 Two approaches to wastewater treatment plant modeling. *Acta Hydrotech.* **24** (40), 45–64.

Li, J. 2011 *Research on the Imbalanced Data Learning*. PhD Thesis, Jilin University, China, 43 pp.

Li, D. C., Liu, C. W. & Hu, S. C. 2010 A learning method for the class imbalance problem with medical data sets. *Comput. Biol. Med.* **40** (5), 509–518.

Lin, H. T. & Lin, C. J. 2003 A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Technical Report, Department of Computer Science and Information Engineering, National Taiwan University website, www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf, visited 24 May 2014.

Liu, Y. H. & Chen, Y. T. 2007 Face recognition using total margin-based adaptive fuzzy support vector machines. *IEEE Trans. Neural Netw.* **18** (1), 178–192.

López, V., Fernández, A., Moreno-Torres, J. G. & Herrera, F. 2012 Analysis of preprocessing vs. cost-sensitive learning for

imbalanced classification. Open problems on intrinsic data characteristics. *Expert Syst. Appl.* **39** (7), 6585–6608.

Mannina, G., Cosenza, A., Vanrolleghem, P. A. & Viviani, G. 2011 A practical protocol for calibration of nutrient removal wastewater treatment models. *J. Hydroinform.* **13** (4), 575–595.

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.

Oh, S. H. 2011 Error back-propagation algorithm for classification of imbalanced data. *Neurocomputing* **74** (6), 1058–1061.

Ranawana, R. & Palade, V. 2006 Optimized precision – A new measure for classifier performance evaluation. *Proc. IEEE Congress on Evolutionary Computation (CEC 2006)*. 16–21 July Vancouver, Canada, pp. 2254–2261.

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

Shamseldin, A. 2010 Artificial neural network model for river flow forecasting in a developing country. *J. Hydroinform.* **12** (1), 22–35.

Wen, C. H. & Vassiliadis, C. A. 1998 Applying hybrid artificial intelligence techniques in wastewater treatment. *Eng. Appl. Artif. Intell.* **11** (6), 685–705.

Widodo, A. & Yang, B. S. 2007 Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors. *Expert Syst. Appl.* **33** (1), 241–250.

Wilson, D. L. 1972 Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Syst. Man Cybernet.* **2** (3), 408–421.

Xiao, J., Xie, L., He, C. & Jiang, X. 2012 Dynamic classifier ensemble model for customer classification with imbalanced class distribution. *Expert Syst. Appl.* **39** (3), 3668–3675.

Yang, Q., Shao, J., Scholz, M. & Plant, C. 2011 Feature selection methods for characterizing and classifying adaptive Sustainable Flood Retention Basins. *Water Res.* **45** (3), 993–1004.

Zhang, Q. & Stanley, S. J. 1999 Real-time water treatment process control with artificial neural networks. *J. Environ. Eng.* **125** (2), 153–160.

Zhang, C., Wu, C. G., Blanzieri, E., Zhou, Y., Wang, Y., Du, W. & Liang, Y. C. 2009 Methods for labeling error detection in microarrays based on the effect of data perturbation on the regression model. *Bioinformatics* **25** (20), 2708–2714.

Zhao, Z. Q. 2009 A novel modular neural network for imbalanced classification problems. *Pattern Recognit. Lett.* **30** (9), 783–788.

First received 9 November 2013; accepted in revised form 5 May 2014. Available online 28 May 2014