

Data Processing for a Water Quality Detection System on Colombian Rio Piedras Basin

Edwin Castillo¹, David Camilo Corrales^{1,2(✉)}, Emmanuel Lasso¹,
Agapito Ledezma², and Juan Carlos Corrales¹

¹ Grupo de Ingeniería Telemática, Universidad del Cauca,
Campus Tulcán, Popayán, Colombia
{efcastillo, dcorrales, eglasso,
jcorral}@unicauca.edu.co

² Departamento de Ciencias de la Computación e Ingeniería, Universidad
Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganés, Spain
davidcamilo.corrales@alumnos.uc3m.es,
ledezma@inf.uc3m.es

Abstract. Freshwater is considered one of the most important of planet's renewable natural resources. In this sense, it is vital to study and evaluate the water quality in rivers and basins. A study area is Rio Piedras Basin, which is the main water supplier source of 9 rural communities in Colombia. Nevertheless, these communities do not make a water quality control. Different research has been conducted to develop water quality detection systems through supervised learning algorithms. However, these research approaches set aside the data processing for improve the outcomes of supervised learning algorithms. This paper presents an improvement of data processing techniques for a water quality detection system based on supervised learning and data quality techniques for Rio Piedras Basin.

Keywords: Water quality data · Lotic ecosystem · Dimensionality reduction · Imbalanced classes · Classifier · DT · ANN · BN · K-NN · SVM · PCA · Boosting and SMOTE

1 Introduction

Freshwater is considered one of the most important of planet's renewable natural resources. In this sense, it is vital to study and evaluate the water quality in lotic ecosystems, which represent water ecosystems in constant motion and in the same direction, such as rivers and basins [1]. A problem occurs in Rio Piedras Basin, which is located on the western slope of the Central Cordillera, west of Popayán (Colombia) and it is the main water supplier source of 9 rural communities: Huacas, Laureles, Canelo, Quintana, San Juan, Santa Teresa, Laguna, San Ignacio, and San Isidrio [2, 3]. Nevertheless, these communities do not make a control water quality.

A significant amount of research has been conducted to develop water quality detection systems, which allow monitoring activities of water quality on different basins [4–7] and lotic ecosystems [8–10] around the world [5, 11–17], through supervised

learning (SL) algorithms. SL tasks predict or classify a new input data from examples (instances), commonly called training data (composed of attributes and a target variable), through algorithms such as decision trees (DT), Bayesian networks (BN), Artificial Neural Networks (ANN), K-Nearest Neighbor (K-NN) and Support Vector Machines (SVM) [18]. However, these research approaches set aside the data quality verification (i.e. redundant attributes, duplicate instances, imbalanced dataset, etc.).

Therefore, an improvement of data processing techniques for water quality detection system for *Rio Piedras Basin*, based on supervised learning techniques, which considered the issues founded on a data quality verification phase. The remainder of this paper is organized as follows: Sect. 2 describes the study area description, the data quality issues addressed and supervised learning algorithms used; Sect. 3 refers the data processing for water quality detection system proposed; Sect. 4 presents results and discussion and Sect. 5 relates conclusions and future work.

2 Background

2.1 Study Area Description

The data used in this study were collected quarterly in the *Rio Piedras Basin*, located in Cauca department, Colombia (source: 76° 31' 10" west of Greenwich and 2° 21' 45" north latitude, mouth of river: 76° 23' 45" west longitude and 2° 25' 40" north latitude), by the Environmental Studies Group (ESG) from the University of Cauca, between 2011 and 2013, taking into account the methodology followed in [19]. Captured samples contain biological (macroinvertebrates) and physicochemical variables, at three points of the basin: Puente Alto, Puente Carretera and Bocatoma Diviso, in different precipitation periods: high (October-November), average (June-July) and low (August-September).

Thus, there were captured 10 physicochemical indicators, 5 biological indicators and 3 precipitation periods. In total, the built dataset consists of 645 records, and 3 values to classify (classes) [17], as set forth in Table 1.

According to the latest work, the three (3) values to classify are denoted by the numbers 1, 2 and 3, which represent a high water quality (very clean water), good (slightly polluted water) and Regular (moderately polluted), respectively.

2.2 Dimensionality Reduction

The dimensionality reduction is the transformation of high-dimensional data in a meaningful representation of smaller dimensions. This reduced representation must have the minimum number of parameters required for expressing the observed data properties [20, 21]. The mentioned task is primarily oriented towards two objectives: instances and attributes reduction techniques.

Attributes Reduction (AR). The attribute reduction decreases the dimension of the attributes within a dataset [21–24]. AR methods are grouped into two categories: attribute selection and extraction; the first looks the best subset of features according to

Table 1. Dataset *Rio Las Piedras* attributes

Category	Attribute	Unit of measurement	Range	Class
<i>Physicochemical indicators</i>	Temperature	°C	13.0 - 17.8	High Water Quality (1)
	Conductivity	µs/cm	35.2 - 89.0	
	Total dissolved solids	mg/L	16.5-42.1	
	Dissolved oxygen	mg/L	7.17-8.23	
	pH	mg/L	6.62-8.17	Good Water Quality (2)
	Ammonium	mg/L	0.01-0.04	
	Nitrates	mg/L	0.01-0.09	
	Nitrites	mg/L	0.01-0.06	
Phosphates	mg/L	0.08-0.24	Good Water Quality (2)	
Turbidity	mg/L	1.0-9.8		
<i>Biological indicators</i>	Class	-	-	Low Water Quality (3)
	Order	-	-	
	Family	-	-	
	Taxon	-	-	
	Number of individuals	-	-	
<i>Precipitation periods</i>	Month	-	-	Low Water Quality (3)
	Year	-	-	
	Sampling point	-	-	

certain criteria (choice of attributes and/or number of attributes to be selected), discarding redundant, inconsistent and irrelevant attributes, while the second transforms the high dimension attributes set in a space of smaller dimension [25, 26].

Noteworthy is the importance of AR mechanisms, in optimizing a dataset. However, when applying techniques for selecting attributes, information loss is generated [22], which is a problem in small datasets, such as the dataset *Rio Las Piedras*. Therefore, this paper chose to use attribute extraction tactics.

At the same time, it was conducted a systematic review of 44 published researches from 2004 to the present, based on the guidelines set forth in [27], taking as search sources: IEEE Xplore (35 items), ScienceDirect (9 items), focused in application domains such as: intruder detection, medicine, biometrics, facial recognition, satellite images classification, among others, obtaining that the algorithm of Principal Component Analysis (PCA) is the most used (33 items), and also overcomes the capabilities of information viewing and understanding than other extraction techniques. Therefore, this algorithm is taken as a starting point to reduce attributes that allow a water quality evaluation task. The explanation of the chosen techniques is presented in Sect. 3.

Instances Reduction. Instances Reduction (IR) decreases the number of irrelevant instances within a dataset [28–30]. Several authors propose classifications of IR methods, as is the case of [31], in which group IR techniques: Noise Filters, Condensation Algorithms and Prototype Algorithms, while in [32] are classified in:

Wrapper and Filter. Furthermore, in [33–35] ensemble methods as Cascading, stacking, Bagging, Boosting, Random Forest are used for IR tasks.

To select the appropriate algorithm for the instances extraction, 34 studies published from 2006 to the present were reviewed, using as search sources: IEEE Xplore (12 items), ScienceDirect (10 items), Springer Link (6 items) and Google scholar (6 items) and focusing the search in application areas such as: intruder detection, security, classifiers building, time series, text recognition, among others. In this review it was found that the most commonly used techniques are Ensemble and Wrapper with 16 and 12 papers respectively, while “Filter” methods are referenced only 6 times.

It is important to indicate that the Wrapper algorithms tend to be over-trained (overfitting) due to frequent use of cross-validation as evaluation technique on a single dataset, which tends to be adjusted to very specific training data features that do not have no causal relationship with the objective function [36]. Moreover, within the Ensemble methods, Boosting is the IR technique most often used [34] and it has a greater capacity of data generalization. Based on the above reasons, this research will take as its starting point the Boosting algorithm for IR tasks.

2.3 Imbalanced Classes

The classes imbalance problem occurs when the number of class label instances is greater (majority or negative class “C−”) on the number of instances that have other labels class (minority or positive class “C+”) [37, 38], and whose unbalance degree can be measured using *IL (Imbalance level) reason* [39–41].

In this scenario, the classifiers have a tendency classification to the majority class, thereby minimizing the classification error and correctly classifying majority class instances detriment of minority class instances.

To solve the imbalance class problem, there have been two main approaches [38, 42, 43]: the external method (at data level) and the internal method (at classification algorithms level). The first consists of achieving a balance between classes by eliminating instances of the majority class (sub-sampling) or the inclusion of instances in the minority class (over-sampling); while the internal method adjusts the classifiers to favor the class minority. In this comparison the external method is the most versatile (because they do not require any change in the algorithm and also can be used in different application domains), the most widely used in the last decade and whose most representative algorithm is SMOTE (Synthetic Minority Over-Sampling Technique), which adds to the minority class examples by creating new instances (which called synthetic data) obtained from an interpolation process.

2.4 Classifiers for Water Quality Detection

For classifiers selection, there were taken 4 researches as a starting point [18, 44–46], in which performing a literature review and theoretically evaluate supervised learning algorithms most commonly used as the case of Decision Trees (DT), Artificial Neural

Networks (ANN), Bayesian Networks (BN), K-Nearest Neighbor (K-NN) and Support Vector Machines (SVM) considering metrics as: accuracy, noise tolerance, ability of explanation, learning speed and classification speed.

3 Mechanism for Detecting Water Quality in *Rio Piedras Basin*

The water quality detection system in *Rio Piedras Basin*, contains various components to perform a pre-processing data and supervised learning algorithms based on data mining tool: KNIME® Analytics Platform, all transparent to the end user. In Fig. 1 the architecture of the proposed mechanism is exposed, which is made up of data processing and classification modules. The first module consists of two components for dimensionality reduction (attributes and instances) to address the redundant values problem and a component that balances the classes. The second module consists of some supervised learning algorithms: DT, ANN, BN, K-NN and SVM.

Below, the proposed mechanism components are briefly described:

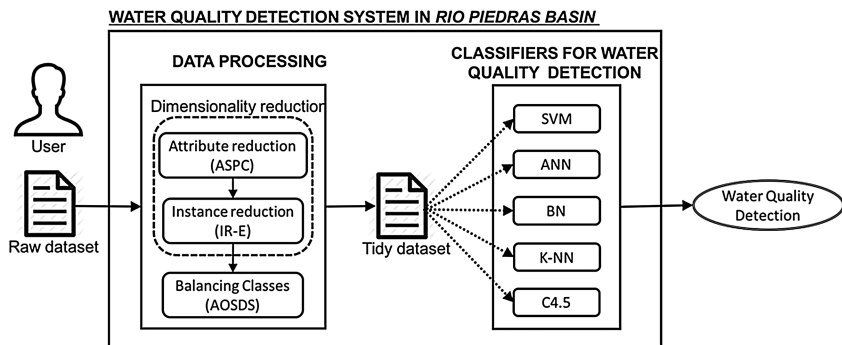


Fig. 1. Mechanism architecture

3.1 Automatic Selection of Principal Components (ASPC)

The Automatic Selection of Principal Component (ASPC) is based on the PCA algorithm definition, and proposes a new approach to Principal Components (PC) selection.

First, PCA is a multivariate technique from data exploratory analysis (statistical), which transforms a large number of attributes, correlated with each other, in a number of attributes smaller uncorrelated (linearly independent or orthogonal), called principal components (PC). The main components are a linear combination of the original attributes, which describe the greatest amount of dataset information and are sorted by the amount of information that they contain (high to low) [47–49].

Currently, there is no definite rule on the exact number of PC to be used. However, to select the optimal PC subset, the authors in [50] exhibit various techniques such as:

B1-Backward, B1-Forward, B2 and B4; obtaining better results with B4 method. B4 intends to use the PC whose accumulated explained variance exceeds 60 % of the information.

On the other hand, ASPC is the mechanism, which consists of building data subsets $X'_{n \times p'}$, for $p' = 1, 2, \dots, p$ and $p = EV$ (PCA determines eigenvalues (EV) and attributes p within the dataset) with p' variables and n observations, from each of the eigenvalues and their eigenvectors.

From here, data subsets are constructed based on the accumulated $\{1\}, \{1+2\}, \{1+2+3\}, \dots, \{1+2+3+m\}$ PC. Subsequently, each data subset is evaluated by means of classification accuracy obtained by some supervised learning techniques: SVM, BN, K-NN, DT and ANN. Finally, it is selected the data subset with better accuracy and thus the number of PC associated with this.

It is worth noting, for the main components analysis should be observed the relationship between PC and the initial attributes from the feature matrix taking into account the correlations sign and magnitude.

3.2 Instances Reduction (IR-E)

The mechanism defined for Instances Reduction (IR-E) takes as a starting point the model proposed in [34], which is based on the Boosting Classifier definition to create a model that selects redundant and irrelevant instances within a dataset and is called BIS (Boosting Instance Selection). This algorithm replaces the classification models for IR algorithms as: Incremental Reduction Optimization Procedure 3 (Drop3), IB3, Iterative Case Filtering (ICF), Modified Selective Subset (MSS), Reduced Nearest Neighbor (RNN), Condensed Nearest Neighbor Rule (CNN).

BIS performs a process M times (M is defined a priori) and, in each iteration, a voting process is executed, which consists of assigning one vote (v_i) to each selected instance (x_i) through a particular IR technique randomly chosen. After M iterations, it is obtained as a result a votes vector V , which records the obtained votes by each instance and, from the latter, a set of thresholds is built $(\emptyset_1, \emptyset_2, \dots, \emptyset_M)$, with the instances which obtained most votes. Now, to get the best threshold it is defined a criteria $J(\emptyset)$, which is represented by an instances subset S_\emptyset belonging to the training set T , so that meet the Eq. 1.

$$\theta : S_\theta = x_i \in T : v_i > \emptyset \quad (1)$$

To asses each $J(\emptyset_i)$ criteria, the K-NN classifier is trained with the subset S_{\emptyset_i} where the quality of each instances subset is selected according to both, the classification performance and the number of removed instances, as seen in the Eq. 2. In fact, any classifier can be used (DT, SVM, BN o ANN); Nevertheless, for this research we will restrict ourselves to use K-NN due to its simplicity, ability to detect wrong results and high speed learning [18, 44–46],

$$J(\theta) = \alpha C + (1 - \alpha)r \quad (2)$$

Where, C symbolizes the classification performance (precision or AUC), r denotes the deleted instances percentage and α is the parameter used to distinguish the importance of each factor.

On the other hand, each threshold assessment involves a K-NN classifier training with the instances subset that represents it, which implies a high degree of complexity of assessment process ($2M + 2$) when number of iterations M is high (the maximum number of possible thresholds is equal to M).

To address the problem, optimal threshold selection called \emptyset_o , from an approach that works on two levels. The first level consists of constructing a subset of thresholds ($\theta_1, \theta_2, \dots, \theta_p$) from the set of possible thresholds ($\emptyset_1, \emptyset_2, \dots, \emptyset_M$) for $p \leq M$, discarding the repeated or equivalent thresholds latter, because these get similar assessments in terms of performance and classification codes. Thus a degree of less than or equal complexity ($2p + 2$) is achieved.

At the second level it seeks to find a threshold value representing both high voting values as low values and thus counteract the drawbacks mentioned above. For this, the arithmetic mean of the thresholds subset resulting from the previous step is determined, which process is represented by the Eq. 3.

$$\emptyset_o = \frac{1}{P} \sum_{i=1}^P \theta_i \quad (3)$$

3.3 Automatic Optimal Synthetic Data Selection (AOSDS)

AOSDS is based on the over-sampling algorithm SMOTE (Synthetic Minority Over-Sampling Technique) definition and proposes a strategy to generate the appropriate number of synthetic data in such a way that minimizes over-training the classifier.

SMOTE currently works with binary classes, a majority and a minority class, denoted as C- and C+ respectively. This algorithm creates synthetic instances or data for the minority class, by interpolating an instance and its closest K neighbors belonging to that class. First, K closest neighbors from the minority class are selected and, subsequently, the synthetic instantiated percentage (denoted by P) is chosen. Then, to generate a new synthetic data, interpolation between the line connecting each minority class instance with any (or all) of their closest K neighbors previously selected is performed. This calculation is made using the Euclidean distance definition and a replacement or overlay (called overlap) function that assigns a value 0 (if both values are equal) or a value 1 (in the case they are different). Since SMOTE only applied to binary classes, this work focused the problem of imbalance dataset Rio Las Piedras on two classes: Class 1 and 3 (C- and C+ respectively) and ignoring the remaining class (class 2). However, it is important to mention that class 2 is only ignored in over-sampling process.

Although this technique generates new instances that allow balancing classes, so far not found a method to indicate the optimal number of instances to be created, since a large number of synthetic instances can over-train the classifier and generate inaccurate results [42, 51, 52].

The proposed strategy involves determining that the original dataset imbalance level (IL) meets a minimum classes threshold unbalance. If the dataset meets the threshold of imbalance, we proceed to apply to the minority class definition SMOTE. For this case, an imbalance level $IL \geq 3.9$ is defined and instances percentage to oversample (P value) equal to 50, 100, 150 and 200 % will be used, thereby preventing the ratio C+ instances exceed those of C-. Furthermore, [53] discloses that to achieve a 200 % of oversampling only required two nearest neighbors, enough for the study case. Therefore, $K = 2$ is taken as the number of nearest neighbors for SMOTE.

Once applied the over-sampling process for a given P , the behavior of each subset generated is evaluated and the data subset that gets the best performance is selected. At the same time, an optimal synthetic data percentage, associated with that dataset, is defined.

3.4 Classifiers to Water Quality Detection

This component consists of classifiers to perform a water quality detection in *Rio Piedras Basin*, which were selected based on the research presented in [18, 44–46] where theoretically evaluate some classifiers: DT, ANN, BN, K-NN and SVM, considering metrics as: precision, noise tolerance, explanatory capacity, learning speed and classification speed. In these researches, is claimed that there is no algorithm that satisfies all evaluation metrics. Also, depending on the dataset used, the algorithm has a different behavior (in metrics of model evaluation such as precision). Therefore, it is necessary to evaluate each of these algorithms in order to obtain a high degree of precision in predictions and easy interpretations. Additionally, it is important to mention that to evaluate the proposed mechanism, cross-validation with $k = 10$ was used.

4 Experimental Results

ASPC, IR-E and AOSDS methods were evaluated individually and collectively, using classifiers as SVM, BN, K-NN, C.4.5 (DT) y ANN, applied to the original and processed datasets. The experimental process can be seen in more detail in the Fig. 2.

4.1 Attributes Reduction

The results of PCA process on the *Rio Las Piedras* dataset are shown in Fig. 3. As shown in Fig. 3(a), 18 principal components (PC) are generated, where first PC represents 21.2 % of the total information, the second PC explains 15.2 % of the original variance, the third explains 11.6 %, and so on until PC 18, which represents 0.1 % of the total information. As discussed above, explained variance data are important to

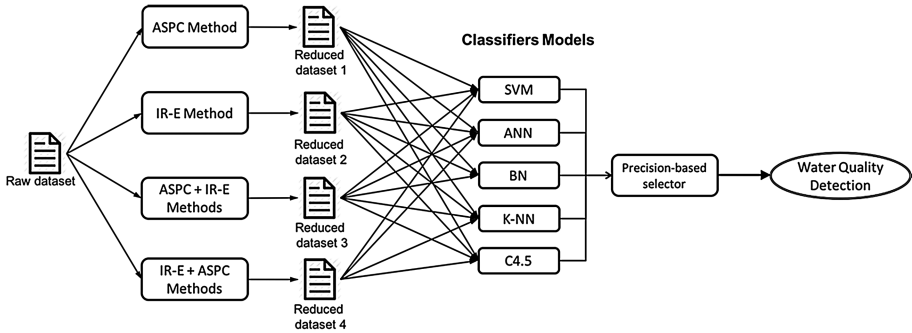
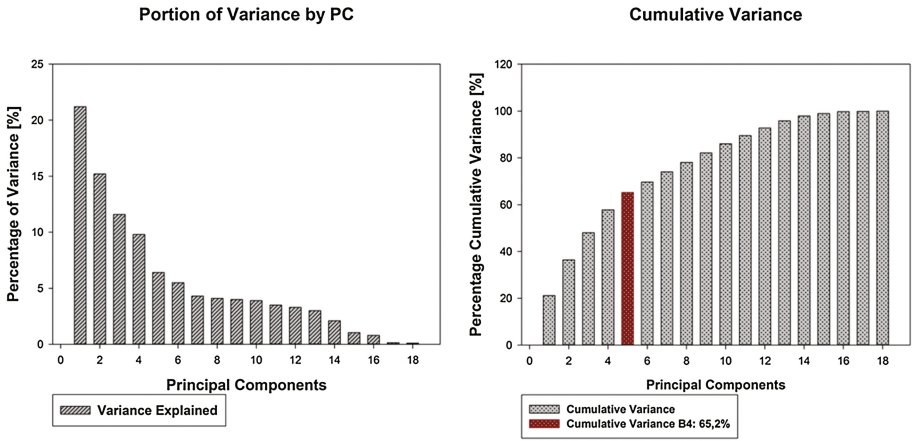


Fig. 2. Experimental process for Dimension Reduction

know the number of PC that will be used in the analysis. If we take B4 methodology as a criterion for components selection, it is considered that the optimal number of PC (reducing the PC amount as much as possible) is 5 components, which variance is 65.2 %. This means that the five (5) first PC represent 65.2 % of the total dataset information. As shown in Fig. 3(b), the other components explain significantly lower percentages comparatively to the first five components.



(a) Variance portion explained by PC (b) Accumulated variance

Fig. 3. PCA results on *Rio Las Piedras* dataset

Once applied the PCA process on the dataset, the ASPC proposed mechanism is applied, which involves as a first step in building a dataset with $\{1\}$, $\{1+2\}$, $\{1+2+3\}$, ..., $\{1+2+3+\dots+18\}$ principal components. Subsequently, the classifiers mentioned above are training with each of these datasets, with a 10-fold cross-validation. Finally, we obtain as a result the data subset that best precision

obtained and the number of PC associated with this dataset. These results are shown in Table 2.

Table 2. ASPC results

Dataset	No. attributes	Information representation	Classifiers precision					Average precision
			BN	SVM	C.4.5	K-NN	ANN	
Not processed	18	100 %	80.2 %	92.8 %	77.1 %	73.2 %	96.1 %	83.8 %
ASPC	5	65.2 %	91 %	91.4 %	91.4 %	86.9 %	89 %	89.9 %

As seen, the ASPC mechanism selects the dataset consisting of the first 5 PC as the best data set, because the latter had the best average precision (89.9 %) among all dataset formed. The reduction of 13 components obtained by the ASPC method led to improved classification performance by approximately 6 %.

In practical terms, the problem initially represented in an 18-dimensional cyber-space has been reduced to a smaller hyperspace (5 dimensions), capturing 65.2 % of the original variance. This result implies greater and easy data interpretability, process that, as mentioned above, is performed by analyzing the relationship between PC and initial attributes.

From these results, it can be assumed that the ASPC method can perform a quite strong selection of training data without deteriorating the capacity of classifiers.

Now, to compare the ASPC method with the B4 selection method in the Table 3, the results obtained are reflected.

Table 3. Comparison between ASPC and B4 methods

ASPC			B4		
No. PC	Information representation	Average precision	No. PC	Information representation	Average precision
5	65.2 %	90.1 %	5	65.2 %	90.1 %

The ASPC mechanism applied on the *Rio Las Piedras* dataset was able to reduce the space to five components. These results corresponded with those obtained with the B4 method, which suggests that the new dataset is constituted with five components and whose information amount is within the accumulated variance confidence interval 60 % – 95 %. Moreover, the mechanism increased both the classification process performance as the model interpretability (fewer attributes).

4.2 Instance Reduction

By applying the IR-E method proposed for instance reduction on a water quality dataset, the results presented in Table 4 were obtained.

Table 4. RI-E results

Dataset	No. attributes	No. instances	No. classes	Classifiers precision					Average precision
				RB	SVM	C.4.5	K-NN	RNA	
Not processed	18	645	3	80.2 %	92.8 %	77.1 %	73.2 %	96.1 %	83.8 %
IR-E	18	336	3	50.6 %	57.1 %	62.8 %	55.2 %	55.2 %	56.1 %

As noted, the IR-E technique achieves to reduce to 309 the instances within the *Rio Piedras* dataset. However, IR-E decreased the classifiers precision in 27.7 %, indicating that important dataset instances were removed, thus losing important information.

4.3 Attributes and Instances Reduction

Under the same context, the Table 5 shows the results product to train classifiers with the newly acquired dataset after applying jointly reduction techniques presented above. As noted, the sequential composition of attribute reduction and subsequent instances reduction is denoted as ASPC + IR-E, while the sequential combination of instances reduction and subsequent attributes reduction is represented as IR-E + ASPC.

Table 5. Attribute and instance reduction methods in conjunction results

Reduction methods	No. attributes	No. instances	Information	Classifier precision					Average precision
				BN	SVM	C.4.5	K-NN	ANN	
ASPC + IR-E	5	493	65.2 %	72.7 %	73.8 %	76 %	72.7 %	74.2 %	73.9 %
IR-E + ASPC	3	225	41.5 %	44.7 %	41 %	60.9 %	65.7 %	59.1 %	54.3 %

It is interesting that the IR-E + ASPC methods sequence reduced in greater quantity both, attributes such as number of instances, in comparison with the technique ASPC + IR-E. The first technique reduced 15 components and 420 instances, while the second reduced 13 components and 152 instances. However, when a classifier is trained with the dataset obtained trough ASPC + IR-E a higher precision (73.9 %) is obtained in comparison when the training process is performed with the dataset obtained trough IR-E + ASPC method (54.3 %). In addition, it may be noted that the dataset reduced by IR-E + ASPC is represented by three characteristics that explain only 41.5 % of the total variance, which is outside the confidence range suggested by the B4 criterion. This means that this method is not appropriate to reduce the *Rio Las Piedras* dataset size.

As a result, we can see that the ASPC + IR-E mechanism is a suitable solution for reducing the water quality dataset size, which allows classifiers provide a similar precision to that obtained with the original dataset (no pre-processing). Accordingly, the IR-E + ASPC method is not appropriate to reduce the dataset size.

4.4 Average Classifiers Training Time

It is noteworthy that another way to evaluate the algorithms performance (efficiency) is by reducing classifiers runtime in dataset processing task. This measure is important when the task of training a classifier for detecting water quality is related to a highly complex scenario (large amounts of data) or a decision support system that requires to define a contingency action in relatively short times.

For this analysis, the arithmetic average time that each classifiers takes to process a dataset is calculated. For this case study, the datasets would be the no pre-processing or original *Rio Las Piedras* dataset and processed or reduced *Rio Las Piedras* dataset. In Fig. 4 these results are summarized.

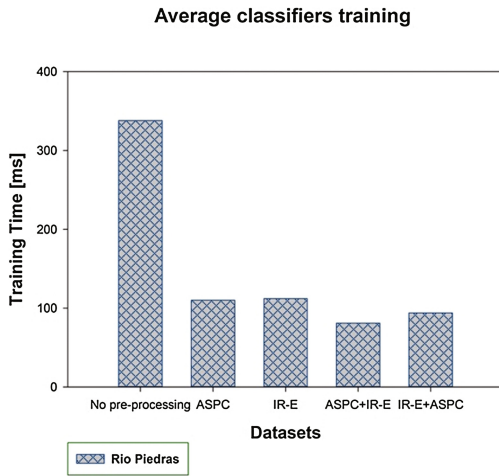


Fig. 4. Average classifiers training

In a first step, it is observed that the ASPC technique, in addition to reduce the number of dataset attributes (13), reduces the classifiers training time in 228 ms in comparison with the training time for the original dataset. This indicates that the existence of redundant features in the dataset influences the performance of classifiers.

Following the same behavior, it is observed that the IR-E technique reduce the dataset in 309 instance and the classifiers training time in 226 ms. The result suggests that by minimizing the amount of redundant instances and noise, it is possible to reduce the time necessary to training the classifiers (computational cost).

Now, in the case of reduction methods combined ASPC + IR-E and IR-E + ASPC applied on the *Rio Las Piedras* dataset, we have that in both cases is greatly reduced the time for training the classifiers (244 ms and 257 ms respectively). Additionally, it is noted that the approach ASPC + IR-E allows a classification process significantly faster than individual methods and even the combined method IR-E + ASPC.

These results are to be expected, because such mechanisms remove both datasets characteristics (instances and attributes), which represent a computational cost reduction in data mining tasks (classification and/or grouping).

As a result, the proposed strategy gets similar precision values to the original dataset and greatly reduces classifiers training time, making this proposal the most appropriate to reduce the dataset size for water quality detection.

4.5 Class Balance

To evaluate the proposed mechanism, a classifiers cross-validation was used with $k = 10$: SVM, BN, K-NN, C.4.5(DT) and ANN. On this approach, an imbalanced dataset *Rio Las Piedras* version (Table 6) and over-sampled datasets with 50 %, 100 %, 150 % and 200 % of synthetic instances (Table 7), was evaluated. As indicated above, the majority and minority class are denoted as C- and C+ respectively.

Table 6. Imbalanced dataset description

No. classes	No. instances	No. instances C-	No. instances C+	IL
3	493	307	54	5.6

The performance of each of these classifiers is obtained in terms of ROC metrics (Receiver Operating Characteristic) and F-Measure (measures in %), considered to evaluate both the original and the over-sampled datasets with different amounts of synthetic instances (50 %, 100 %, 150 % and 200 %). The results are shown in Table 7.

Table 7. Classifiers performance behavior over C- and C+ classes

Class/Dataset	SVM		ANN		C.4.5		BN		K-NN	
	F-M	ROC	F-M	ROC	F-M	ROC	F-M	ROC	F-M	ROC
Original										
C-	87.4 %	77.8 %	86.8 %	78 %	88 %	75.6 %	86.6 %	76.5 %	80 %	75.8 %
C+	0	66.3 %	0	62 %	0	60.8 %	0	60 %	9 %	51.5 %
SMOTE-50										
C-	84 %	74.4 %	84.8 %	76 %	84.7 %	76 %	83 %	73.6 %	78 %	72.8 %
C+	0	60 %	0	62 %	18.6 %	66 %	0	62.5 %	43.6 %	60.8 %
SMOTE-100										
C-	81 %	71.7 %	80.8 %	75 %	84.6 %	81 %	80 %	73 %	75 %	71 %
C+	0	60 %	0	67.7 %	51.6 %	79 %	0	64.8 %	41 %	64 %
SMOTE-150										
C-	78 %	69 %	76 %	64 %	84.5 %	82 %	72 %	73.5 %	74 %	72.5 %
C+	0	61.4 %	21 %	68 %	59.5 %	79.8 %	29 %	68.5 %	47.7 %	66.5 %
SMOTE-200										
C-	69 %	75.7 %	72 %	72.6 %	79 %	80.7 %	67.7 %	72.5 %	73 %	72 %
C+	0	72 %	21.5	68.6	60 %	81 %	40.6	69 %	53 %	67.7 %

As indicated in the above table, to train the C.4.5 classifier with the original *Rio las Piedras* dataset ($IL = 5.6$), instances belonging to the negative class tend to be correctly classified (M-F = 88 % y ROC = 75.6 %), while those belonging to the positive class tend to be classified incorrectly (M-F = 0 y ROC = 60.8 %). This is because the classification methods tend to favor the negative class.

Similarly, increasing the number of positive class instances with synthetic data, the classification performance of the latter improves, obtaining the best results when synthetic data at 150 % and 200 % are generated, with F-M = 59.5 %, and ROC = 79.8 % and F-M = 60 %, ROC = 81 % respectively. Although there is no significant difference in these results, and considering that what is sought in class balance process is that the classification process can correctly predict C+ instances without affecting significantly the C- instances detections, the best classification results are obtained when C+ is balanced with 150 % of synthetic instances. This oversampling level improvement performance metrics F-Measure and ROC in C+ class at 59.5 % and 19 % respectively, without affecting the C- class detection, moreover, affects only in 3.5 % the F-Measure and increasing the ROC value by 6.4 %.

Now, the K-NN measures shown in Table 7 exposes clearly the same behavior of the previous classifier (C.4.5), where the C+ class is oversampling at 150 % of synthetic instances and this allows the classifier to discriminate it better and detect it correctly (F-Measure and ROC metrics increase their performance to 38.7 % and 15 % respectively), at the expense of 6 % of precision and 3.3 % of ROC area reduction for C- class. Meanwhile, when C+ class is oversampled with a sampling level of 200 %, F-Measure and ROC metrics improves at 5.3 % and 1.2 % with respect to the previous process, further reducing the performance of C- classification.

On the other hand, the SVM classifier got for each C+ class a value of F-Measure = 0. Otherwise it occurs in the C- class, where good results are obtained. From here, we can say that SVM classifier is very sensitive to the class imbalance problem, because this algorithm has its separation hyper-plane very close to the minority class, resulting in a low or null performance for examples of this class, compared to those of the majority class [54].

Like the BN, the ANN are more difficult to profit in the metrics when the dataset is oversampled with a percentage of less than 150 % synthetic instances. In fact, like the previous cases, the best results for all cases are obtained when the dataset is oversampled with this value (150 %), and can be seen as the best balance between the number of true and false positives.

From these results, we can see how in spite of balance the dataset trough SMOTE technique, the percentage of correctly classified instances by SVM, BN and ANN algorithms does not significantly increase, and is even lower than the results obtained when classifiers are trained directly with imbalanced dataset (original). Therefore, the fact of using synthetic samples generated trough SMOTE represents a gain in the discriminant capacity of classes, but not equal in all classifiers.

In this vein, the experimental results showed that oversample the imbalanced dataset with levels of 150 % and 200 % improves the detection of classifiers in C+ class without significantly impairing the detection of the C- class. In addition, it is observed that by using a level of imbalance or another, the same results are obtained

approximately. However, to create too many synthetic instances produces overtraining, which affects the ability to detect the minority class.

From the foregoing, it is considered that the best level of synthetic instances oversampling is 150 %, and the SMOTE algorithm is considered as a practical method for generating synthetic instances. However, we must be especially careful in the number of generated instances, as it can make poor results in C- class prediction.

Based on the considerations submitted, in the Table 8 balanced dataset (with 150 % of synthetic instances) is described.

As shown in the above table, to oversample the dataset with 150 % of synthetic instances increase in 54 the examples of *Rio Las Piedras* dataset.

4.6 Classification Module

This section presents the classifiers evaluation (DT, ANN, BN, K-NN, SVM) and results analysis, applied on the dataset described in the Table 1 and on the processed dataset after applying the dimensionality reduction and balancing classes approach (Table 8).

Table 8. Overview of processed dataset

Dataset	No. attributes	No. instances	No. classes
<i>Rio Las Piedras</i>	5	547	3

As mentioned above, the performance of each of these classifiers is obtained in terms of some metrics as: precision, recall, F-Measures and ROC, considered to evaluate both the original as the processed dataset (proposed mechanism). The results are presented in Fig. 5.

In the case of the original dataset, the ANN and C.4.5 algorithms obtained the best results among all evaluated classifiers, incorrectly classifying a smaller instances proportion, compared to the other methods, which can be checked with the obtained precision (C.4.5 = 83 % and RNA = 76 %, compared to SVM = 62.6 %, K-NN = 70.1 % and BN = 65 %).

Additionally, it is important to mention that the C.4.5 classifier gets the highest proportion of true positives than other techniques because it has a higher recall (83 %), while other algorithms obtained a recall value not more than 70 % for the case of SVM, BN and K-NN. Regarding to F-Measure, the C.4.5 and ANN followed the same behavior, as they obtain the best values (83.3 % and 76.3 % respectively) among all the evaluated techniques. These results suggest that C.4.5 and ANN supervised algorithms have a best behavior for working with this dataset.

Similarly, the five supervised learning algorithms are trained with the processed dataset (Table 8), as shown in Fig. 5. This graph shows that for all classification models, the number of incorrectly classified instances was reduced, except for C.4.5, where remained approximately constant. This behavior is reflected in the increased

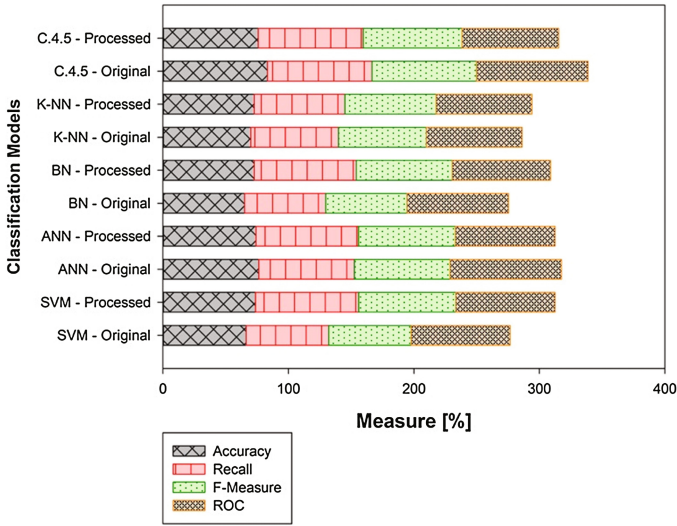


Fig. 5. Result of classifiers evaluation: SVM, ANN, K-NN, BN and C.4.5 applied to *Rio Las Piedras* original and processed datasets

precision of classifiers: BN (72.7 %), SVM (73.8 %) and K- NN (72.7 %) with the exception of ANN which failed to increase its accuracy (74.2 %).

However, the true positives proportion in relation to false positives is high, since the recall values exceed 72 % (SVM = 82.2 %, ANN = 81.7 %, BN = 81.3 %, K-NN = 72.4 % and C.4.5 = 83.6 %). The mentioned results let us to identify that the 5 evaluated models have a good confidence since the number of false positives is low and the number of relevant instances classified is high, as can be contrasted through F-Measure values (SVM = 77.3 %, ANN = 77.2 %, BN = 76.6 %, K-NN = 72.6 % and C.4.5 = 78.8 %).

5 Conclusions and Future Works

In the previous sections, mechanisms for datasets dimensionality reduction and water quality detection were proposed. From the results achieved, it can be inferred that the proposed mechanism (ASPC + IR-E) is an appropriate solution for pre-processing water quality datasets, in order to reduce their size, which allows classifiers to provide a similar precision to that obtained with the original dataset (no pre-processing). Similarly, the execution times to classifiers training tasks were reduced.

Additionally, the evaluated classification algorithms present good results. However, ANN and C.4.5 algorithms obtained the highest precision values, classifying fewer instances wrong and also kept the same behavior in all experiments. Therefore, these supervised learning models are chosen as the base classifiers to detect water quality in *Rio Piedras Basin*, taking into account the dataset characteristics associated with this domain.

As future work, it is intended to deploy the proposed mechanisms in a production environment; address other dataset problems, such as missing values and outliers; and use a similar methodology to generate a water quality prediction, in order to get the ability to construct an early warning system for lotic ecosystems.

Acknowledgments. The authors are grateful to the University of Cauca and its Telematics Engineering Group (GIT) for the technical support and the Environmental Studies Group (GEA) for providing the dataset, Control Learning Systems Optimization Group (CAOS) of the Carlos III University of Madrid, KNIME® Analytics Platform, AgroCloud project of the RIC-CLISA Program, Ministerio de Economía y Competitividad de España (Proyecto TRA2011-29454-C03-03. i-Support: Sistema Inteligente Basado en Agentes de Soporte al Conductor) and Colciencias (Colombia) for PhD scholarship granted to MsC. David Camilo Corrales.

References

1. Álvarez, L.F., Arango, M.C., Arango, G.A., Torres, O.E., de Jesús Monsalve, A.: Calidad Del Agua De Las Quebradas La Cristalina Y La Risaralda, San Luis, Antioquia. In: EIA, pp. 121–141, Julio 2008
2. Marchant, C., Mergili, M., Borsdorf, A.: Agricultura Ecológica y Estrategias de Adaptación al Cambio Climático en la Cuenca del Río Piedras. Cuenca Río Las Piedras (2012)
3. Acosta, M., Devereux, T.: Manual de las medidas de adaptación al cambio climático practicadas por los campesinos de Asocampo de la cuenca Río Las Piedras, Cauca, Colombia: Un resumen visual de las medidas de adaptación local frente al cambio climático y el trabajo y la investigación en campo. Centro Internacional de Agricultura Tropical CIAT 2013
4. Dang, J., Huo, A.-D., Song, J.-X., Chen, X.H., Mao, H.-R.: Simulation modeling for water governance in basins based on surface water and groundwater. *Agric. Water Manage.* (2016)
5. Sun, W., Liao, H.: Forecasting and evaluating water quality of chao lake based on an improved decision tree method. *Procedia Environ. Sci.* **2**, 970–979 (2010)
6. Lek, S., Cheng, L., Lek-Ang, S., Li, Z.: Predicting fish assemblages and diversity in shallow lakes in the Yangtze River basin. *Limnologica* **42**, 127–136 (2012)
7. Zhang, W., Wang, Y., Engel, B.A., Peng, H., Theller, L., Shi, Y., Hu, S.: A fast mobile early warning system for water quality emergency risk in ungauged river basins. *Environ. Model Softw.* **73**, 76–89 (2015)
8. Yan, J., Tan, G., Gao, C., Yang, S.: Prediction of water quality time series data based on least squares support vector machine. *Procedia Eng.* **31**, 1194–1199 (2012)
9. Basant, N., Gupta, S., Singha, K.P.: Support vector machines in water quality management. *Anal. Chim. Acta* **703**, 152–162 (2011)
10. Liong, S.-Y., Tkalich, P., Palani, S.: An ANN application for water quality forecasting. *Mar. Pollut. Bull.* **56**, 1586–1597 (2008)
11. Xu, J., Liao, Y., Wang, W.: A method of water quality assessment based on biomonitoring and multiclass support vector machine. *Procedia Environ. Sci.* **10**, 451–457 (2011)
12. Sophatsathit, P., Areechakul, S., Lursinsap, C.: Integration of unsupervised and supervised neural networks to predict dissolved oxygen concentration in canals. *Ecol. Model.* **261–262**, 1–7 (2013)
13. Tai, H., Liua, S., Ding, Q., Li, D., Xu, L., Wei, Y.: A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. *Math. Comput. Model.* **58**, 458–465 (2012)

14. Bucak, I.O., Karlik, B.: Detection of drinking water quality using CMAC based artificial neural networks. *Ekoloji Dergisi* **20**, 75–81 (2011)
15. Park, Y.-S., Bae, M.-J.: Biological early warning system based on the responses of aquatic organisms to disturbances: a review. *Sci. Total Environ.* **466–467**, 635–649 (2014)
16. Gupta, S., Singh, K.P.: Artificial intelligence based modeling for predicting the disinfection by-products in water. *Chemometr. Intell. Lab. Syst.* **114**, 122–131 (2012)
17. Gonzales, W.F., Castillo, E.F., Corrales, D.C., López, I.D., Hoyos, M.G., Figueroa, A., Corrales, J.C.: Water quality warnings based on cluster analysis in Colombian rivers basins. *Sistemas y Telemática (S&T)* **13**, 9–26 (2015)
18. Corrales, J.C., Corrales, D.C., Figueroa-Casas, A.: Towards detecting crop diseases and pest by supervised learning. *Ing. Univ.* **19**, 207–228 (2015)
19. Pérez, G.R.: Bioindicación de la Calidad del Agua en Colombia: Propuesta Para el Uso del Método BMWP Col, Primera ed. vol. 1. Universidad de Antioquia (2003)
20. Fukunaga, K.: Introduction to Statistical Pattern Recognition. School of Electrical Engineering-Purdue University-West Lafa yet te, Indiana
21. Inza, I., Larrañaga, P., Saeys, Y.: A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**, 2507–2517 (2007)
22. Khalil, K., Nasreen, S., Khalid, S.: A survey of feature selection and feature extraction techniques in machine learning. In: Science and Information Conference (SAI), 27–29 August 2014, pp. 372–378 (2014)
23. Wang, X., Paliwal, K.K.: Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern Recogn.* **36**, 2429–2439 (2002)
24. Deepa, T., Ladha, L.: Feature selection methods and algorithms. *Int. J. Comput. Sci. Eng. (IJCSSE)* **3**, 1787–1797 (2011)
25. Popescu, M.C., Sasu, L.M.: Feature extraction, feature selection and machine learning for image classification: a case study. In: IEEE (2014)
26. Paliwal, K.K.: Dimensionality reduction of the enhanced feature set for the HMM-based speech recognizer. *Digit. Sig. Process.* **2**, 157–173 (1992)
27. Kitchenham, B.: Procedures for performing systematic reviews. Joint Technical report, July 2004
28. Ahmad, S.S.S., Pedrycz, W.: Feature and instance selection via cooperative PSO. In: IEEE, 9–12 October 2011, pp. 2127–2132 (2011)
29. Tsai, C.-F., Chang, C.-W.: SVOIS: support vector oriented instance selection for text classification. *Inf. Syst.* **38**, 1070–1083 (2013)
30. Chan, Z.-Y., Ke, S.-W., Tsaia, C.-F.: Evolutionary instance selection for text classification. *J. Syst. Softw.* **90**, 104–113 (2014)
31. Jankowski, N., Grochowski, M.: Comparison of instances selection algorithms I. Algorithms survey. In: Rutkowski, L., Siekmann, J.H., Tadeusiewicz, R., Zadeh, L.A. (eds.) ICAISC 2004. LNCS (LNAI), vol. 3070, pp. 598–603. Springer, Heidelberg (2004)
32. Ariel Carrasco-Ochoa, J., Arturo Olvera-López, J., Francisco Martínez-Trinidad, J., Kittler, J.: A review of instance selection methods. *Artif. Intell. Rev.* **34**, 133–143 (2010)
33. Blachnik, M.: Ensembles of instance selection methods based on feature subset. *Procedia Comput. Sci.* **35**, 388–396 (2014)
34. García-Pedrajas, N., De Haro-García, A.: Boosting instance selection algorithms. *Knowl.-Based Syst.* **67**, 342–360 (2014)
35. Blachnik, M., Kordos, M.: Bagging of instance selection algorithms. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2014, Part II. LNCS, vol. 8468, pp. 40–51. Springer, Heidelberg (2014)

36. Jordan, M.I., Karp, R.M., Xing, E.P.: Feature selection for high-dimensional genomic microarray data. In: ICML 2001 Proceedings of the Eighteenth International Conference on Machine Learning, pp. 601–608 (2001)
37. Hong, X., Gao, M., Chen, S., Harris, C.J.: A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems. *Neurocomputing* **74**, 3456–3466 (2011)
38. Fernández, A., García, S., Herrera, F.: Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems. *Appl. Soft Comput.* **9**, 304–1314 (2009)
39. Elrahman, S.M.A., Abraham, A.: A review of class imbalance problem. *J. Netw. Innovative Comput.* **1**, 332–340 (2013)
40. Satyasree, K.P.N.V., Murthy, J.V.R.: An exhaustive literature review on class imbalance problem. *Int. J. Emerg. Trends Technol. Comput. Sci. (IJETTCS)* **2**, 109–118 (2013)
41. Verbiest, N., Ramentol, E., Cornelis, C., Herrera, F.: Improving SMOTE with fuzzy rough prototype selection to detect noise in imbalanced classification data. In: Pavón, J., Duque-Méndez, N.D., Fuentes-Fernández, R. (eds.) *IBERAMIA 2012*. LNCS, vol. 7637, pp. 169–178. Springer, Heidelberg (2012)
42. Cooper, E.W., Nguyen, H.M., Kamei, K.: Borderline over-sampling for imbalanced data classification. In: *Fifth International Workshop on Computational Intelligence & Applications* (2009)
43. Kamel, M.S., Sun, Y., Wong, A.K.C., Wang, Y.: Cost-sensitive boosting for classification of imbalanced data. *Pattern Recogn.* **40**, 3358–3378 (2007)
44. Bhavsar, H., Ganatra, A.: A comparative study of training algorithms for supervised machine learning. *Int. J. Soft Comput. Eng. (IJSCE)* **2**, 74–81 (2012)
45. Kotsiantis, S.B.: Supervised machine learning: a review of classification techniques. *Informatica* **31**, 249–268 (2007)
46. Zaharakis, I.D., Pintelas, P.E., Kotsiantis, S.B.: Machine learning: a review of classification and combining techniques. *Artif. Intell. Rev.* **26**, 159–190 (2006). Springer Science
47. Wu, C.-M., Zhang, Y., Luo, Y.: Facial expression feature extraction using hybrid PCA and LBP. *J. China Univ. Posts Telecommun.* **20**, 120–124 (2013). ScienceDirect
48. Xu, D., Wang, Y.: An automated feature extraction and emboli detection system based on the PCA and fuzzy sets. *Comput. Biol. Med.* **37**, 861–871 (2007)
49. Xiao, B.: Principal component analysis for feature extraction of image sequence. In: *International Conference on Computer and Communication Technologies in Agriculture Engineering*, 12–13 June 2010, vol. 1, pp. 250–253 (2010)
50. King, J.R., Jackson, D.A.: Variable selection in large environmental data sets using principal components analysis. *Environmetrics* **10**, 67–77 (1999)
51. Makond, B., Wang, K.-J., Chen, K.-H., Wang, K.-M.: A hybrid classifier combining SMOTE with PSO to estimate 5-yearsurvivability of breast cancer patients. *Appl. Soft Comput.* **20**, 15–24 (2014)
52. Sicilia, M.Á., Riquelme, J.C.: SMOTE-I: mejora del algoritmo SMOTE para balanceo de clases minoritarias. *Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos* **3**, 73–80 (2009)
53. Bowyer, K.W., Chawla, N.V., Hall, L.O., Philip Kegelmeyer, W.: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
54. He, H., Ghodsi, A.: Rare class classification by support vector machine. In: *2010 20th International Conference on Pattern Recognition (ICPR)*, 23–26 August 2010, pp. 548–551 (2010)