DESIGN AND IMPLEMENTATION OF INTELLIGENT SEAWATER AUTOMATIC ON-LINE MONITORING SYSTEM BASED ON BIG DATA

L.Y. LI[†], J. YANG[‡], Y. LEI[§], K.H XIONG[⊥], W.H. CHEN[†], K.F. LIN[‡], Y.T. CHEN[‡], Y.Y. YU[‡], T.T. CHEN[‡], X.C. LI[‡], and S.YAN[† ‡*]

† *Branch of Fujian Normal University, Fuqing, China.*
‡ *Oceanology Institute of East China Sea, Fujian, China.*
§ *Environmental Protection Science Research Institute, Ningde, China.*
⊥ *Environmental Monitoring Station, Longyan, China.*
383987929@qq.com

***Abstract*— Based on large data analysis method and automatic detection technology, this paper designs a test system, which can realize intelligent online monitoring of seawater. Based on the theory of large data, the data preprocessing method of large data is applied by relying on the information transmitted by integrated sensors. Using data cleaning, data integration, data conversion and data reduction technology, a large number of data collected by marine monitoring devices are processed accurately. An automatic seawater monitoring system is designed on a software platform. Finally, combined with the experimental data of a certain sea area, the test results are analyzed, which proves the feasibility and effectiveness of the designed seawater online monitoring system. It has achieved the effect of seawater environmental analysis and early warning.**

***Keywords*— Big data; data preprocessing, on-line monitoring.**

I. INTRODUCTION

With the development of economy, more and more countries in the world need energy. In the process of exploitation and utilization of marine resources, the marine environment has also been destroyed. Therefore, the research and development of seawater environmental monitoring technology is particularly important. In the process of developing and utilizing marine resources, marine environmental monitoring technology provides technical support for people. Once the marine environment is damaged, fisheries and aquaculture will suffer huge losses. Therefore, obtaining accurate and reliable data parameters and accurately predicting the marine environment are of great significance to the marine economic industry.

Since the late 1960s, many countries have begun to study marine environmental monitoring technology, and have carried out many marine environmental investigation plans and monitoring projects in the international secret areas (Lin *et al.*, 2018). During this period, marine environmental monitoring technology has developed rapidly, and all the major countries in the world have a variety of physical and chemical sensors (Chocarro-Ruiz *et al.*, 2018). The United States, Russia, Canada, Norway, Japan and other countries have developed a variety of marine pollution on-line monitoring sensors. Using advanced marine technology, the United States, Russia, Norway and other countries have established a complete water pollution monitoring buoy system to monitor water quality (Lofthu *et al.*, 2018). In addition, various water quality monitoring sensors have been added to the ocean buoys. He has the US, the EB52 buoy, the Russian AK300 buoy and the Tobey buoy in Norway. Improvement of marine environmental monitoring (Zhang *et al.*, 2018a). At present, the mature pollution monitoring sensors include dissolved oxygen sensors, turbidity, pH value and salinity. Other sensors that indicate the degree of pollution of the marine environment are still being researched and developed.

The technical level of marine environmental monitoring sensors is the concrete embodiment of the level of marine monitoring automation (Izquierdo *et al.*, 2018). At present, the basic situation of marine ecological environment monitoring is: current, water temperature, salinity, atmospheric pressure, dissolved oxygen sensor technology has matured, its accuracy and stability has reached a fairly high level, and chemical biosensor has not passed (Seyyedi *et al.*, 2018). At present, the marine ecological environment monitoring technology is still mainly based on sampling and sample analysis. The development of marine ecological environment monitoring technology based on biological principles is relatively slow.

At present, seawater environmental monitoring instruments have been developed along two ways: one is direct measurement method, that is, using sensors to directly measure pollution parameters underwater; the other is the "micro laboratory" method (Hussain *et al.*, 2014), which simplifies the sampling and analysis process to miniaturization and solves the current technical problems. The latter has been commercialized and has become the mainstream of water quality monitoring instruments such as YS 16820, YS 16920 and Min 2I Sonde of HYDOLAB.

Under the background of big data, this paper designs an intelligent seawater automatic on-line monitoring system. Using the data preprocessing method of large data,

a large number of data collected by the marine monitoring device are processed accurately (Sahoo *et al.*, 2018). In addition, based on C, VB and LabView platform, the software of lower computer, upper computer and monitoring center are designed. Finally, the feasibility and validity of the seawater on-line monitoring system are proved by the experimental data of seawater on-line monitoring.

The rest of this paper is arranged as follows. The background knowledge needed to design an on-line seawater monitoring system is introduced in Part 2. An on-line seawater monitoring system was designed in 3. In 4, the effectiveness of the system is proved by experimental data. Finally, the research results of this paper are summarized, and some prospects for the future work are also given.

II. RESEARCH BACKGROUND

In the big data environment, there are some problems in the raw data from heterogeneous systems:

Lack of standardized criterion. The original data is obtained from each practical application system. Because the data of each application system lacks a unified standard definition, and the data structure is also quite different, the data of each system is inconsistent and cannot be used directly.

Because of the intrinsic complexity of data types and organizational patterns, complex relationships, and uneven quality, data perception, expression, understanding and computation are facing enormous challenges (Zhang *et al.*, 2018b). Therefore, data preprocessing is a very important data preparation work before data mining, and is one of the key links of knowledge discovery in database (KDD). On the one hand, it can ensure the correctness and validity of the data mining, on the other hand, through the adjustment of data format and content, make the data more in line with the needs of mining. The pretreatment methods for dirty data are as follows: cleaning, integration, transformation and reduction.

A. Data cleaning

There are redundancy, error, inconsistency and other noise data in the detection data. Using various cleaning techniques, a "clean" consistent data set is formed. As shown in Fig.1. Data cleaning techniques include eliminating duplicate data, filling missing data, and eliminating noise data (Canito *et al.*, 2018). After analyzing the origin and existing form of "dirty data", we can make full use of new technical means and methods to clean "dirty data" and convert "dirty data" into data that meets the requirements of data quality or application. The United States began to study data cleaning technology very early. With the development of information industry and commerce, data cleaning technology has been further developed.

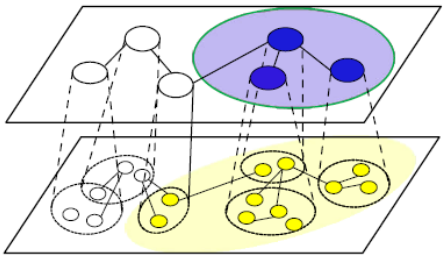

Figure 1. Data cleaning.

Missing data cleaning. Improving missing data is another important issue in the field of data cleaning. As shown in Fig. 2, in the real world, the content of a data set is incomplete due to various reasons, such as misoperation of manual input, confidentiality of some information or unreliable data sources. If the incomplete data set is large, once deleted a number of records, because the remaining data set is small, making the model construction does not have universality and representativeness (Rathore *et al.*, 2016), cannot be trusted, reliability greatly reduced. In addition, the data set deviation caused by deleting incomplete data makes the classification and clustering model of data mining incline seriously, which affects the final mining results and produces significant decision-making misleading.

Statistical methods were used to fill in missing values. Analyze the data set, get the statistical information of the dataset, and fill in the missing values with numerical information. Fill in missing values by classification and clustering (Gepp *et al.*, 2018). Classification is to construct a classifier (such as classification function or classification model) by inputting training sample data sets on the basis of existing class labels. Using the record similarity between complete records and missing records, the maximum possible complete data model is established by calculating the maximum similarity and combining with the relevant technology of machine learning (Alharthi *et al.*, 2017). Clustering is to seek the similarity between classes without considering the label of classes. The purpose of clustering is to construct a small representative data set based on a large amount of data aggregation, and to further analyze and study the set.

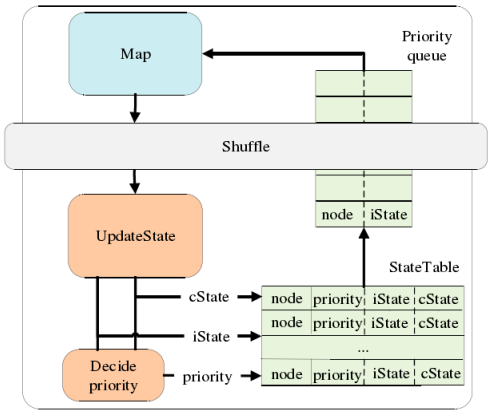

Figure 2. Missing data cleaning.

Noise data processing. Before data mining (Shin, 2016), it is often assumed that there is no data interference in the dataset. However, due to various reasons, a large number of noise data, i.e. outliers, are produced in the process of data collection and collation.

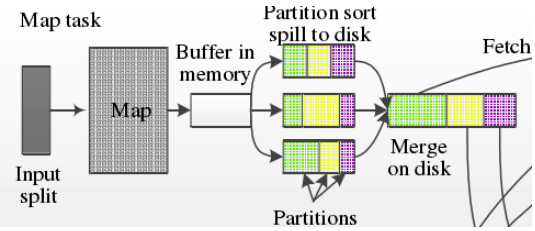

Figure 3. Noise data processing.

In order to improve the speed and accuracy of data mining, it is necessary to remove duplicate records in data sets. If two or more instances represent the same entity (Banica *et al.*, 2016), then duplicate records. In order to find duplicate instances, it is common practice to compare each instance with other instances to find the same instance. By calculating the similarity of the attributes of the records and considering the different weights of each attribute, the similarity of the records can be obtained after weighted average. If the similarity between two records exceeds a certain threshold, the two records are considered to be matched, otherwise (Hu *et al.*, 2015), the two records are considered to point to different entities. Another similarity computation algorithm is based on the basic nearest neighbor sorting algorithm. The core idea is to reduce the number of comparisons of records, move a fixed-size window on the keyword-sorted data set, and detect the records in the window to determine whether they are similar, so as to determine the duplicate records.

B. Data Integration

Data integration is to merge heterogeneous data in multi-file or multi-database running environment to solve semantic ambiguity. It mainly involves data selection, data conflict and inconsistent data processing, as shown in Fig.4.

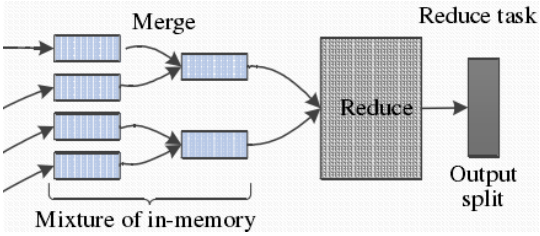

Figure 4. Data Integration

C. Data transformation.

Data transformation is to find the characteristic representation of data, reduce the number of valid variables or find invariants of data by dimension transformation or transformation (Wu *et al.*, 2013), including normalization, switching and projection operations. Data transformation is the transformation of data into a form suitable for various mining patterns, according to the subsequent use of data mining algorithms, to decide which data transformation method to use.

III. DESIGN OF SEAWATER ON-LINE MONITORING SYSTEM

Developing the ocean, utilizing the ocean resources and developing the ocean economy is an important part of the national economic development and the forward position of the sustainable development strategy (Triguero *et al.*, 2015). However, with the development of the ocean, although it has brought enormous economic benefits to people, it also brings environmental problems. At present, marine pollution sources are mainly divided into two parts: terrigenous pollution and oil production. In addition, organic matter, radioactive substances and nutrients are flowing into the ocean in large quantities, and industrial heat pollution is increasing.

All kinds of marine organisms have a suitable range of water temperature, pH, EH and dissolved oxygen for their growth (Galar *et al.*, 2012). If they go beyond this range, they will affect their survival and even death. Seawater temperature and pH are suitable for the growth and reproduction of organisms, but also enrich plankton resources. At the same time, sea water temperature and pH are also important factors causing red tide.

In summary, temperature, PH, EH and dissolved oxygen are selected as the main parameters in the marine chemical/physical monitoring system. In addition, the measurement of dissolved hydrogen sulfide, carbon dioxide concentration and pressure parameters are added.

A. Data preprocessing

The processing of massive data is a great challenge to the existing technology. The emergence of large data makes people obtain unprecedented large-scale samples when dealing with computational problems (Gao *et al.*, 2011), but at the same time they have to face more complex data objects. As an important data preparation before data analysis and mining, data preprocessing can ensure the accuracy and validity of data mining results.

In order to improve the efficiency of the on-line monitoring system, a large number of data collected by the on-line monitoring device are preprocessed. A large number of data processing steps are as follows:

Step 1: Ignore incomplete data. By directly removing attributes or instances, (Gallego, 2016) the incomplete data is ignored. This method is often used to achieve data cleaning in the case of small data sets and less incomplete data. This method is often used as a default method because of its high execution efficiency.

Step 2: Use the instance reduction algorithm to reduce the size of the dataset. Using CNN (condensed nearest neighbor) algorithm, ENN (edited nearest neighbor) algorithm, ICF (iterative case filtering) algorithm and DROP (decremental reduction by ordered projections) algorithm to reduce the amount of data without reducing the quality of knowledge acquisition. Using LVQ (learning vector quantization) algorithm, the data size is greatly reduced by removing or generating new instances (Sotoca *et al.*, 2010). For dimensionality reduction

(dimensionality reduction) of high-dimensional data. The technologies involved include feature selection (feature selection) and spatial transformation (space transformations). The core of dimension reduction is to reduce the number of random variables or attributes.

Step 3: Use discretization technology. MDLP (minimum description length principle) algorithm and ChiMerge, CAIM (class-attribute interdependence maximization) algorithm in the mapping of discernibility matrix are used to reduce the number of given continuous attribute values. Before discretization, the scale of discrete data should be predicted first (Pérez-Ortiz *et al.*, 2015), and then the sequential data should be sorted, and then several splitting points should be designated to divide the data into multiple intervals. All continuous data falling in the same interval are mapped to the same discrete data by a unified mapping method.

Step 4: imbalanced learning. SMOTE (synthetic minority oversampling technique) algorithm is used to preprocess the data to avoid uneven distribution of types. Using undersampling method to create subsets of original data sets for data mining by sampling, most instances are removed as far as possible.

B. Online monitoring system

1. Program design

In order to debug the low-power data acquisition device conveniently, we developed the upper computer software of the low-power data acquisition device under the L bVIE environment based on the communication protocol. The main interface of the software is shown in Fig.5. The main functions of the host computer software are as follows:

- Data table: displaying raw data or calculated data of each channel sensor.
- Dynamic temperature line: dynamic display of temperature lines displayed by temperature sensors.
- Current temperature indication: showing the temperature value measured by the current temperature sensor.
- Switch and communication control: control sensor work switch.
- File storage: storing files in the specified file (7) program function: advanced command and option function.

Function Description:

CMMAND dialogue. Click on the CMMAND button and a dialog box appears with five options to send some advanced instructions in the dialog box.

Clicking the option button will pop up a dialog box with 3 options. In this dialog box, you can set the program. As shown in Fig.6.

2. Software design of monitoring center

The real-time data from the lower computer of the data acquisition system is received through the data center. Additionally, to realize the remote control of the lower computer through the monitoring center (Marín *et al.*, 2016), it is necessary to compile the monitoring center software of the upper computer of the marine environment monitoring system, according to the communication protocol. The monitoring center software of the system is compiled under the VB development environment (Angiulli *et al.*, 2007). WinSock network programming interface is used to monitor the data transmitted by GPRS. The data is received and connected with the database, and the data is saved to the database. Access database is selected and VB is used as the development tool of the database to realize the management of monitoring information and enhance the security, activity and efficiency of the database management.

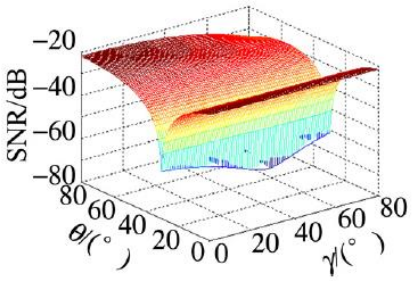

Figure 5. Dialog box under CMMAND option dialogue.

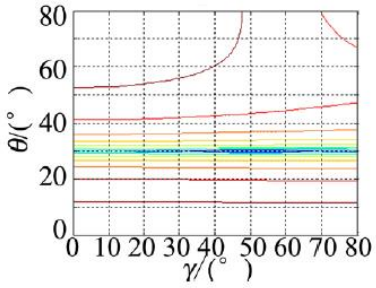

Figure 6. Dialog box under OPTION.

IV. TEST RESULTS AND ANALYSIS

The concentration of H2S is 0.17 nmol/LN 0.26 nmol/L as shown in Fig. 8. The variation of H2S concentration is gentle in line. The Eh value is 116.5 mV to 125.5 mV, which is higher than that in normal still water (Eh in still water is generally between 100 and 110), possibly due to the strong current at the test site.

In Fig. 8, the upper line represents the sea pressure temperature, the lower line represents the temperature, and the pressure responds to the height of the water level, that is, the change of the tide level. It can be clearly seen from the chart that the change of water level in the middle is due to the influence of strong typhoon during the monitoring period (Prati *et al.*, 2015), the pressure fluctuation range becomes larger, and then returns to normal. The following line reflects the change in temperature, and the trend of change during typhoons is quite obvious.

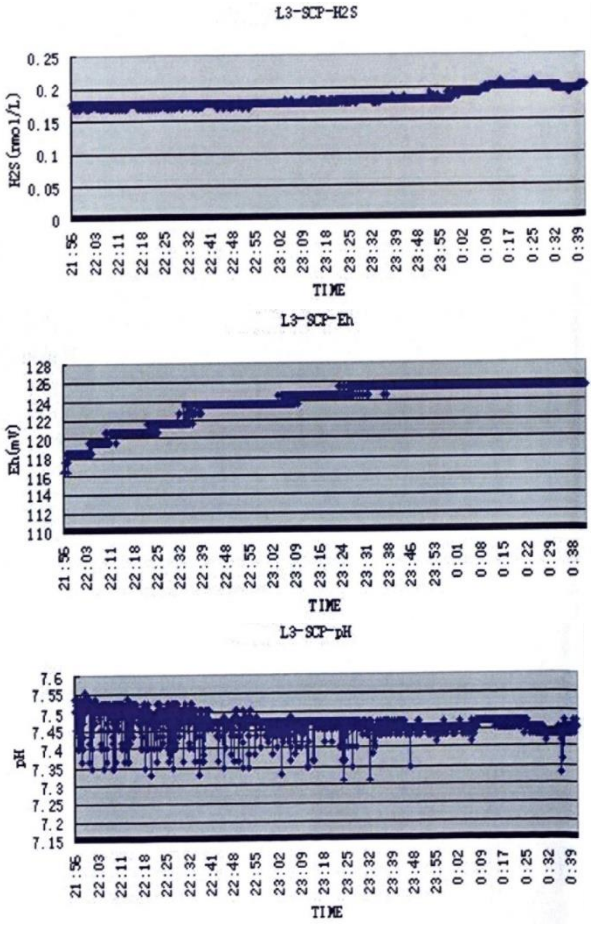

Figure 7. Changes of H2S, Eh and pH values.

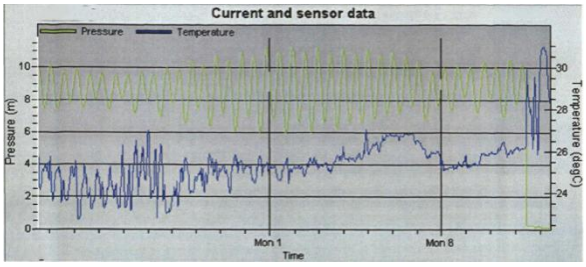

Figure 8. Seawater pressure and temperature variation line.

V. CONCLUSIONS

An intelligent seawater automatic on-line monitoring system is designed in this paper (Brown *et al.*, 2018). This paper analyzes the main tasks in the pretreatment process, summarizes several common processing methods for various types of "dirty data", and focuses on the common algorithms in the process of data cleaning, integration, transformation and reduction. Through various pretreatment methods, redundant data is cleared, error data is corrected, incomplete data is perfected, and necessary data is selected for integration, which makes data information concise, data format consistent and data storage centralized (Bacardit *et al.*, 2012). Data mining on the most accurate and reliable minimum data set greatly reduces the cost of system mining and improves the accuracy, effectiveness, and practicability of knowledge discovery. Finally, the experimental data are obtained in a sea area, and the analyzed results prove the feasibility and effectiveness of the designed seawater online monitoring system.

REFERENCES

Alharthi, A., V. Krotov and M. Bowman, "Addressing barriers to big data", *Business Horizons*, **60(3)**, 285-292 (2017).

Angiulli, F. and G. Folino, "Distributed Nearest Neighbor-Based Condensation of Very Large Data Sets", *IEEE Transactions on Knowledge & Data Engineering*, **19(12)**, 1593-1606 (2007).

Bacardit, J., P. Widera and A. Márquez-Chamorro, "Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features", *Bioinformatics*, **28(19)**, 2441-2448 (2012).

Banica, L. and A. Hagiu, "Using big data analytics to improve decision-making in apparel supply chains", *Information Systems for the Fashion & Apparel Industry*, **20(1)**, 63-95 (2016).

Brown, T., S. Du, H. Eruslu and F.J. Sayas, "Analysis of models for viscoelastic wave propagation", *Applied Mathematics and Nonlinear Sciences*, **3(1)**, 55-96 (2018).

Canito, J., P. Ramos and S. Moro, "Unfolding the relations between companies and technologies under the Big Data umbrella", *Computers in Industry*, **99**, 1-8 (2018).

Chocarro-Ruiz, B., S. Herranz and A. Fernández, "Interferometric nanoimmunosensor for label-free and real-time monitoring of Irgarol 1051 in seawater", *Biosensors & Bioelectronics*, **117**, 47-52 (2018).

Galar, M., A. Fernandez, and E. Barrenechea, "A review on ensembles for the Class Imbalance Problem, Bagging-, Boosting-, and Hybrid-Based Approaches", *IEEE Transactions on Systems Man & Cybernetics Part C. Applications & Reviews*, **42(4)**, 463-484 (2012).

Gallego, F. B, "On problems of Topological Dynamics in non-autonomous discrete systems", *Applied Mathematics and Nonlinear Sciences*, **1(2)**, 391-404 (2016).

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

Gepp, A., M. K. Linnenluecke, and T. J. O'Neill, "Big Data techniques in auditing research and practice. Current trends and future opportunities", *Social Science Electronic Publishing*, **40**,102-115 (2018).

Hu, H., Y. Wen and Y. Gao, "Toward an SDN-enabled big data platform for social TV analytics", *IEEE Network*, **29(5)**, 43-49 (2015).

Hussain, A., E. Cambria, B. Schuller and N. Howard. "Affective neural networks and cognitive learning

systems for big data analysis", *Neural Networks*, **58 (52)**, 1-3 (2014).

Izquierdo, C. G., A. Garcia-Benadí and P. Corredera, "Traceable sea water temperature measurements performed by optical fibers", *Measurement*, **127**, 124-133 (2018).

Lin, M., X. Hu and D. Pan, "Determination of iron in seawater, from the laboratory to in situ, measurements", *Talanta*, **188**, 135-144 (2018).

Lofthus, S., I. K. Almås and P. Evans, "Biodegradation in seawater of PAH and alkylphenols from produced water of a North Sea platform", *Chemosphere*, **206**, 465-473 (2018).

Pérez-Ortiz, M., P.A. Gutiérrez and C. Hervás-Martínez, "Graph-based approaches for over-sampling in the context of ordinal regression", *IEEE Transactions on Knowledge & Data Engineering*, **27(5)**, 1233-1245 (2015).

Prati, R. C., G. E. A. P. A. Batista and D.F. Silva, "Class imbalance revisited, a new experimental setup to assess the performance of treatment methods", *Knowledge & Information Systems*, **45(1)**, 1-24 (2015).

Rathore, M.M., A. Ahmad, and A. Paul, "Urban planning and building smart cities based on the Internet of Things using Big Data analytics", *Computer Networks*, **101(C)**, 63-80 (2016).

Sahoo, P. K., S. K. Mohapatra and S.L. Wu, "SLA based healthcare big data analysis and computing in cloud network", *Journal of Parallel & Distributed Computing*, **119**, 121-135 (2018).

Seyyedi, M., S. Tagliaferri and J. Abatzis, "An integrated experimental approach to quantify the oil recovery potential of seawater and low-salinity seawater injection in North Sea chalk oil reservoirs", *Fuel*, **232**, 267-278 (2018).

Shin, D. H., "Demystifying big data. Anatomy of big data developmental process", *Telecommunications Policy*, **40(9)**, 837-854 (2016).

Sotoca, J. M. and F. Pla, "Supervised feature selection by clustering using conditional mutual information-based distances", *Pattern Recognition*, **43(6)**, 2068-2081 (2010).

Triguero, I., D. Peralta and J. Bacardit, "MRPR, A MapReduce solution for prototype reduction in big data classification", *Neurocomputing*, **150 A,** 331-345 (2015).

Wu, X., X. Zhu and G.Q. Wu, "Data Mining with Big Data", *IEEE Transactions on Knowledge & Data Engineering*, **26(1)**, 97-107 (2013).

Zhang, Z. M., H. H. Zhang and Y.W. Zou, "Distribution and ecotoxicological state of phthalate esters in the sea-surface microlayer, seawater and sediment of the Bohai Sea and the Yellow Sea", *Environmental Pollution*, **240**, 235-247 (2018).

Zhang, Q., L. T. Yang and Z. Chen, "A survey on deep learning for big data", *Information Fusion*, **42**,146-157 (2018).

Received: December 15th 2017
Accepted: June 30th 2018
Recommended by Guest Editor
Juan Luis García Guirao