Journal of Information Hiding and Multimedia Signal Processing ©2016 ISSN 2073-4212
Ubiquitous International Volume **7**, Number **4**, July **2016**

A Semi-supervised Approach for Water Quality Detection Based on IoT Network

Ye Yuan

Beijing Laboratory of Advanced Information Networks, Beijing, 100124
College of Electronic Information & Control Engineering
Beijing University of Technology
No.100, Pingleyuan, Chaoyang District, Beijing 100124, China
yuanye91@emails.bjut.edu.cn

Ke-Bin Jia*

Beijing Laboratory of Advanced Information Networks, Beijing, 100124
College of Electronic Information & Control Engineering
Beijing University of Technology
No.100, Pingleyuan, Chaoyang District, Beijing 100124, China
*Corresponding author
kebinj@bjut.edu.cn

Received November, 2015; revised May, 2016

Abstract. *Water quality detection is very important for monitoring water sources and main canal, which is beneficial to offer strategies for the management of water quality and environment. According to the practical distribution and data characteristic, this paper proposes a semi-supervised detection method of water quality based on a sparse autoencoder network. In the proposed approach, an IoT-based distributed structure is implemented to execute data interaction, and a representation model is firstly learned via a sparse autoencoder trained by unlabeled water monitoring data acquired from 8 physical reservoirs, then a softmax classifier is trained using a small set of labeled classification data based on the China Surface Water Environmental Quality Standard (GB3838-2002) expressed by the sparse autoencoder. The combined model is finally used to evaluate the water quality. Compared Experimental results with the traditional methods and actual measure results show that the proposed method has high robustness and accuracy for water quality assessment, and has a good prospect of practical applications.*
Keywords: Water quality detection; Sparse autoencoder; Softmax; Semi-supervised learning; IoT.

1. **Introduction.** The uneven distribution of water resources leads to water shortage in China, and serious water scarcity in certain areas. The South-to-North Water Diversion Project is an effective national solution for alleviating the drought of north and northwest in China. Water quality detection is an important means for monitoring water sources and main canal, which is beneficial to offer strategies for the management of water quality and environment. Many of these water monitoring programs generally provide real-time data acquisition and contains mess data with complex construction, which requires the intelligence use of resources using the detection method.

There have been some existing methods used to evaluate water quality grades, such as single factor method [1], grey correlation method [2, 3], multivariate statistical method [4], fuzzy mathematics method [5, 6, 7], etc. However, since these methods are affected

by evaluation mode and subjective parameters for some specified occasions, the results sometimes are not reliability for generalized situation. Moreover, these methods need to be improved to represent the nonlinear relationship between water quality assessment grades and evaluation factors.

Other methods for evaluating water quality are based on machine learning such as support vector machine (SVM) [8, 9], back-propagation neural network (BP-NN) [10, 11], radial basis function neural network (RBF-NN) [12, 13], which have a better adaptability. However, the performance of these supervised learning methods is dependent on the quality of the training set, which need a variety of balanced data with correct label. Unfortunately, in actual situation, it is easy to get a large amount of unlabeled water quality data instead of labeled one. Recently, sparse autoencoder, a deep learning algorithm [14], has been shown great performance in many applications, which conforms to a semi-supervised learning mechanism [15, 16] to construct a better feature representation learned from the unlabeled data before using supervised classifier.

In this paper, we thus propose a semi-supervised water quality detection method based on a sparse autoencoder network. In the proposed approach, an IOT-based distributed structure is implemented to execute data interaction, which consists of a few slave servers and a master server. In each slave server, a representation model is firstly learned via a sparse autoencoder trained by unlabeled water monitoring data automatically acquired from actual IoT network [17]. Then a softmax classifier is trained using a small set of labeled classification data expressed by the sparse autoencoder. The combined model is finally used to evaluate the water quality and send messages to the master server using web technologies, which the results are more effective and objective.

2. **IoT Based structure design.** A distributed structure is designed according to the physical multimedia environment model. The structure based on the internet of things (IoT) framework is shown in Fig.1. The perception layer is composed of various sensors from different manufacturer which collect many properties of water quality such as PH, temperature, ammonia etc.; the network layer transfers those monitoring data to the corresponding slave server through serial port, USB port or internet. When a new data is arrived, combined with location information and other attributes from slave database, the proposed detection method based on proposed method is triggered to distributed detect water quality, and inform to the master server; the application layer provides a B/S system implementing visualization result and other functions for users based on the master server and database.

3. **Method.** Deep learning, a new branch of machine learning, is a set of algorithms that attempt to build a neural network imitating human mind pattern for analysis and learning, which is close to artificial intelligence (AI) and conforms to a semi-supervised learning mechanism. The concept was first proposed by Geoffrey Hinton in Science in 2006 [14], which is an unsupervised greedy layer-wise training algorithm based on deep networks. By constructing a better representation and learning model with deep architecture, deep learning is used to learn more high-order hidden features to increase the accuracy of prediction or classification.

There are mainly two prime steps for training a deep network with good parameters: pre-training and fine-tuning. A greedy layer-wise unsupervised training method is used for pre-training with the given unlabeled data. The parameters from each layer is trained individually while fixing the other parameters of the model, which is the significant difference with traditional neural network method. Then, fine-tuning using supervised learning method can be used to improve the results by changing the parameters from all layers

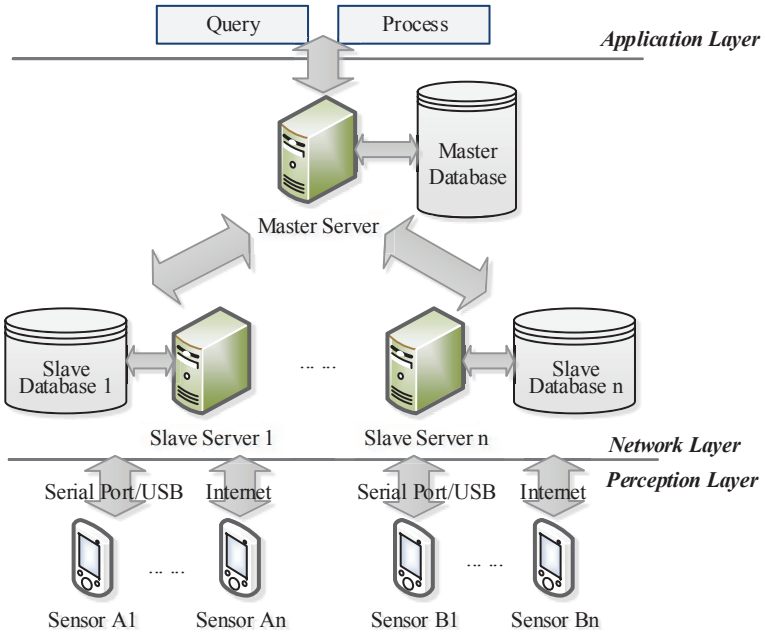

FIGURE 1. The IoT-based structure of water quality detection

simultaneously using labeled data. Therefore, comparing with the traditional shallow learning, the deep architecture transforms the features to a new dimension, which enhances the importance of features to express more deep information from data.

3.1. **Sparse autoencoder.** The basic autoencoder neural network is an unsupervised feature learning algorithm which can be represented as a three-hierarchical directed graph with a visible layer, a hidden layer and a reconstruction layer. It attempts to let the output vector $\hat{x}$ approximates the input vector x by learning an encoder and a decoder. According to the structure of basic sparse autoencoder shown in Fig.2, the sparse autoencoder is parameterized by $(W, b) = (W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)})$ where $W^{(l)}$ is the weight matrix associated with the connection between layer l and layer $l+1$, and $b^{(l)}$ denotes the bias in layer l. For a given single training example x, the hypothesis function of the output based on forward-propagation is:

$$\hat{x} = h_{W,b}(x) = f(W^{(2)} f(W^{(1)} x + b^{(1)}) + b^{(2)}) \tag{1}$$

Here $f(\cdot)$ is an activation function which is defined by sigmoid logistic function as $f(z) = 1/(1 + \exp(-z))$.

To get more sparse features, the sparse constraint is used in the autoencoder hidden layer, and thus it is called a sparse autoencoder. Formally, given the unlabeled training dataset $\{x^{(i)}, i = 1, 2, ..., m\}$ where $x^{(i)} \in \mathbb{R}^n$, by minimizing the discrepancy between the input $x^{(i)}$ and the output $\hat{x}^{(i)}$, the cost function of sparse autoencoder is defined as:

$$J_{sparse}(W, b) = \frac{1}{2m} \sum_{i=1}^{n} \|x^{(i)} - h_{W,b}(x^{(i)})\|^2 + \frac{\lambda}{2} \|W\|^2 + \beta \sum_{j=1}^{s} KL(\rho \parallel \hat{\rho}_j) \tag{2}$$

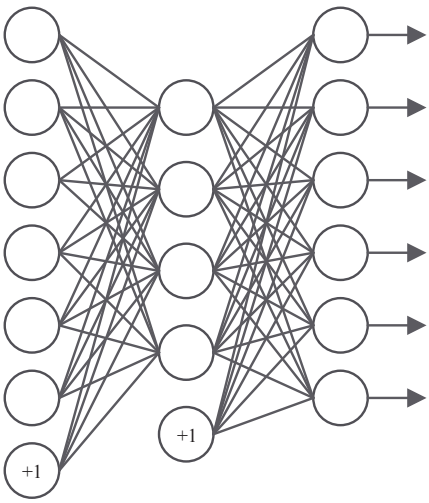

FIGURE 2. The structure of sparse autoencoder

The first term is an average sum-of-squares error term which describes the discrepancy over the entire training data. The second term is a regularization term that tends to decrease the magnitude of the weights, and helps prevent overfitting. The third term is a sparsity penalty term where ρ is called sparsity parameter and $\hat{\rho}$ is called the average activation of hidden unit j (averaged over the training set). $KL(\rho \parallel \hat{\rho}_j)$ denotes the Kullback-Leibler (KL) divergence between a Bernoulli random variable with mean ρ and a Bernoulli random variable with mean $\hat{\rho}_j$, which is defined as:

$$KL(\rho \parallel \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j} \tag{3}$$

The optimization problem can be solved by neural activation forward passing and error back propagation. The gradient of the parameters in sparse autoencoder can be computed accurately so that the L-BFGS algorithm [18], an advanced optimization method, can be used to converge the problem.

3.2. **Softmax classifier with sparse autoencoder.** The softmax regression is a supervised learning algorithm which generalizes logistic regression, a binary classification model, to a multi-class classification model. Formally, given the labeled training dataset $\{(x^{(1)}, y^{(1)}), ..., (x^{(m)}, y^{(m)})\}$, where $y^{(i)} \in \{1, 2, ..., k\}$, the hypothesis of softmax takes the form:

$$h_\theta(x^{(i)}) = \begin{bmatrix} p(y^{(i)} = 1 \mid x^{(i)}; \theta) \\ \vdots \\ p(y^{(i)} = k \mid x^{(i)}; \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} \exp\left(\theta_j^T x^{(i)}\right)} \begin{bmatrix} \exp\left(\theta_1^T x^{(i)}\right) \\ \vdots \\ \exp\left(\theta_k^T x^{(i)}\right) \end{bmatrix} \tag{4}$$

Here $\theta = [\theta_1^T, ..., \theta_k^T]^T$ denotes the parameters of the model, and the term $1/\sum_{j=1}^{k} \exp\left(\theta_j^T x^{(i)}\right)$ is for normalization.

Thus the cost function of softmax is described as:

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\} \log \frac{\exp(\theta_l^T x^{(i)})}{\sum_{l=1}^{k} \exp(\theta_l^T x^{(i)})}] + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n} \theta_{ij}^2 \qquad (5)$$

Here $1\{\cdot\}$ is the indicator function and the he second term is the regularization term which penalizes large values of the parameters. Noticed that in the first term $\frac{\exp(\theta_l^T x^{(i)})}{\sum_{l=1}^{k} \exp \theta_l^T x^{(i)})} = p(y^{(i)} = 1 \mid x^{(i)}; \theta)$, which means the cost function of softmax is similar to logistic except the sum over the k. With the weight decay term, the cost function is strictly convex, and the algorithms such as gradient descent and L-BFGS can be used to converge the global minimum.

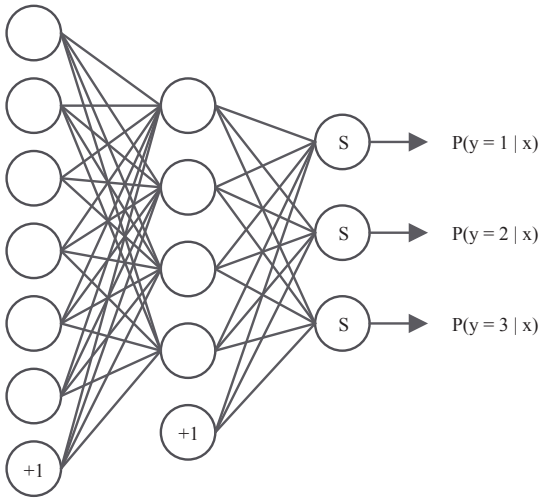

FIGURE 3. The proposed water detection method framework

In this paper, the "decoding" layer of the sparse autoencoder is discarded and the last hidden layer is linked to the softmax classifier, which the proposed water detection method framework is shown in Fig.3. According to the semi-supervised learning mechanism, during the pre-training step, the unlabeled data is used to train a sparse autoencoder for each layer individually. After that, the combined model with softmax classifier is trained using BP algorithm as supervised learning. The final model is used to evaluate the water quality, which the results are more objective and effective.

4. **Experiments and Results.**

4.1. **Historical data organization.** The historical data was obtained from the Water Quality Monitoring Platform (WQMP) [19] for the Central Line Project of South-to-North Water Diversion Project, which contains 8 physical locations which include TaoCha, BaShang, TaiZiShan, ShiJiaWan, ShenDingHe, ZhangYing, and DanJiangKou with different attribution from February 2011 to December 2012. The data is mainly organized in three groups: basic information data, water monitoring data and water quality classification data. The basic information data is consisted of location attribute information and data acquisition equipment information. Some of them need to be considered for

preparation which is important for the accuracy of the model. Noticed that from the given dataset, the water monitoring data is mostly unlabeled and the water quality classification data is manually labeled from some of the monitoring data according to the China Surface Water Environmental Quality Standard (GB3838-2002).

4.2. **Data preprocessing.**

4.2.1. *Feature extraction.* Before training the model, it is necessary to filter the data to eliminate disturbance and smoothen out noise. Hence, some data is removed by using database query which includes duplicate items, zero items and meaningless items. According to the actual situation, we filter the data based on the basic information from the dataset mentioned above for example. 14 representative features of water quality are extracted after filtering which include chemical oxygen demand (COD), dissolved oxygen (DO), total phosphorus (TP), 5-day biochemical oxygen demand (BOD5), NH3-N, NO3-N, oil, chlorophyll, PH, electrical conductivity (EC), turbidity, CL, total coliform (TColi) and temperature (T). After data filtering, only 7,932 unlabeled records and 60 unbalanced labeled classification records from 8 physical reservoirs remained. Water quality is classified into five levels according to the China Surface Water Environmental Quality Standard (GB3838-2002) which is published by Chinese government. Some of the evaluated factors standard grades are shown in Table 1.

Table 1. Surface water environment quality standard (Part) (Mg/L)

Index	Grade				
	I	II	III	IV	V
DO	≥7.5	6	5	3	2
COD	≤15	15	20	30	40
BOD5	≤3	3	4	6	10
NH3-N	≤0.15	0.5	1.0	1.5	2.0
TP	≤0.02	0.1	0.2	0.3	0.4
Oil	≤0.05	0.05	0.05	0.5	1.0

4.2.2. *Data normalization.* The data needs to be represented using a normalized scale for classifier. For the selected m samples $\{x^{(i)}, i = 1, 2, ..., m\}$ where $x^{(i)} \in \mathbb{R}^n$ is a vector and $n = 14$ denotes the 14 extracted features. In order to improve the processing speed and accuracy, each feature dataset data is normalized as follows:

$$x_j^{(i)} = \frac{x_j^{(i)} - \mu_j}{\max(x_j) - \min(x_j)} \tag{6}$$

Here $x_j = (x_j^{(1)}, x_j^{(2)}, ..., x_j^{(m)})^T$ denotes the vector for a specific feature. $\max(x_j)$ and $\min(x_j)$ represent the minimum and maximum value in the vector x_j respectively. μ_j is the average value of x_j.

4.3. **Model training and testing.** The offline method is used to train the model using MATLAB, which will not bring additional computing burden to water classification assessment module, and can satisfy real-time requirement. Due to the unbalance problem of the given classification records, which means it is too few to construct a good softmax classifier with complete training data for all water quality grades, some training samples are linearly inserted and generated on the basis of random method with the uniform distribution between any two assessment grades. Thus according to the 60 unbalanced labeled classification records, 100 training samples are generated for each water quality assessment grade. Finally we have 500 labeled training samples and 7,932 unlabeled records in all.

According to the size of feature vector and the number of output classification, we set a 14-10-5 3-layer deep network. In the proposed approach, we firstly use the normalized unlabeled water monitoring data to unsupervised learn a representation model by sparse autoencoder. After the pre-training step, we use the labeled classification training data to supervised learn the combined network with softmax classifier. Noticed that the 10-fold cross validation is used during the fine-tuning step.

Table 2. Comparison of five classification algorithms

Value	Method				
	Softmax	BP-NN	RBF-NN	SVM	Proposed
Accuracy	88.7%	94.1%	95.6%	97.2%	98.8%

Experimental results show the validity and robustness of the proposed method in compare with traditional Softmax, 3-layer BP-Neural Network, 3-layer RBF-Neural Network and SVM (C=350, g=0.4). The results are reported in Table 2, compared with other supervised algorithm, the proposed semi-supervised leaning method improves the accuracy and reliability. The results illustrate that the proposed method can construct a better data representation due to the pre-training step using unlabeled data which the other supervised baselines cannot be fully used. On the other hand, in this situation, the conditional restriction make it difficult to train a complete model using supervised method with such unbalanced labeled data.

4.4. **Water quality assessment in action.** In order to further verify the effects of the proposed method, the IoT-based distributed water quality assessment function module is implemented by JAVA, which is actually deployed in the Water Quality Monitoring Platform (WQMP). The business functions including alarm mechanism are implemented by Ajax-SSH2 framework [20] based on B/S technologies. Fig.4 shows the data visualization interface of B/S system as an example.

5. **Conclusion.** The use of artificial intelligence facilitates decision making and enables advance monitoring and control. According to the practical situation, this paper proposes a novel semi-supervised method for water quality detection and develops the application of

FIGURE 4. Data visualization interface of B/S system

deep learning. The experimental results demonstrate that the proposed method achieves better recognition rate and meanwhile is fit for future intelligent multimedia management and demand response applications. This work can be extended to a large scope in the future.

Acknowledgment. This paper is supported by the Key Project of Beijing Municipal Education Commission under Grant, Project for the Innovation Team of Beijing, the National Natural Science Foundation of China under Grant No.81370038, the Beijing Natural Science Foundation under Grant No.7142012, the Beijing Nova Program under Grant No.Z141101001814107.

REFERENCES

[1] Y. Bao, S. Liu, G. Zhang, P. Huang, and D. Hou, Methods for water quality abnormity assessment based on single factor, *Reliability Engineering and System Safety*, vol.140, pp.99-106, 2015.

[2] X. Meng, G. Fan, X. Cao, J. He, and J. Qu, Research on a multi-index comprehensive evaluation method for surface water quality assessment, *Advanced Materials Research*, vol.1010-1012, pp.321-324, 2014.

[3] H. Wong and B. Q, Hu. Application of improved extension evaluation method to water quality evaluation, *Journal of Hydrology*, vol.509, pp.539-548, 2014.

[4] T. C. Ogwueleka, Use of multivariate statistical techniques for the evaluation of temporal and spatial variations in water quality of the Kaduna River, Nigeria, *Environmental Monitoring and Assessment*, vol.87, pp.137, 2015.

[5] S. Mike, M. Francisco, and G. Luis, A fuzzy multicriteria categorization of water scarcity in complex water resources systems, *Water Resources Managemen*, vol.29, pp.521-539, 2015.

[6] W. Wang, D. Xu, K. Chau, and G. Lei, Assessment of river water quality based on theory of variable fuzzy sets and fuzzy binary comparison method, *Water Resour. Manage*, vol.28, pp.4183-4200, 2014.

[7] E. Aghaarabi, F. Aminravan, R. Sadiq, M. Hoorfar, M. J. Rodriguez, and H. Najjaran, Comparative study of fuzzy evidential reasoning and fuzzy rule-based approaches: an illustration for water quality assessment in distribution networks, *Stochastic Environmental Research and Risk Assessment*, vol.28 pp.655-679, 2014.

[8] W. Li, M. Yang, Z. Liang, Y. Zhu, W. Mao, and J. Shi, Assessment for surface water quality in Lake Taihu Tiaoxi River Basin China based on support vector machine, *Stochastic Environmental Research and Risk Assessment*, vol.27, pp.1861-1870, 2013.

[9] M. Fereshteh and A. Shahab. A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification, *Water Resources Management*, vol.28, pp.4095-4111, 2014.

[10] M. Xue, A novel water quality assessment method based on combination BP neural network model and fuzzy system, *Journal of Computers*, vol.8, pp.1587-1593, 2013.

[11] E. Farmaki, N. Thomaidis, and C. Efstathiou, Artificial neural networks in water analysis: Theory and applications, *International Journal of Environmental Analytical Chemistry*, vol.90, pp.85-105, 2010.

[12] Q. Luan, C. Zhu, Surface water quality evaluation using BP and RBF neural network, *Journal of Software*, vol.6, pp.2528-2534, 2011.

[13] D. Wang, S. Li, and X. Zhou, Assessment method of raw water quality based on PSO-RBF neural network model and its application, *Dongnan Daxue Xuebao.* vol.41, pp.1019-1023, 2011.

[14] G. E. Hinton, and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, *Science*, Vol.313, pp.504 C 507, 2006.

[15] B. Jiang and K. Jia, Semi-supervised facial expression recognition algorithm on the condition of multi-pose, *Journal of Information Hiding and Multimedia Signal Processing*, vol.4, no.3, pp.138-146, 2013.

[16] H. Yuan, A semi-supervised human action recognition algorithm based on skeleton feature, *Journal of Information Hiding and Multimedia Signal Processing*, vol.6, no.1, pp.175-182, 2015.

[17] F. Chang and D. Chen, Future Classroom with the Internet of Things A Service-Oriented Framework, *Journal of Information Hiding and Multimedia Signal Processing*, vol.6, no.5, pp.869-881, 2015.

[18] C. Zhu , R. H. Byrd and J. Nocedal, L-BFGS-B: Algorithm 778: LBFGS-B: Fortran routines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, vol.23, pp.550-560, 1997.

[19] Z. Pang, K. Jia, Designing and accomplishing a multiple water quality monitoring system based on SVM, in *9th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, IEEE, pp.121-124, 2013.

[20] Y. Yuan, K. Jia. Design and implementation of operation energy management system based on AJAX-SSH2, in *1st Euro-China Conference on Intelligent Data Analysis and Applications (ECC 2014)*, IEEE, pp.355-364, 2014.