Research on Detection Algorithm of Multi dimensional Outlier Based on Weighted Entropy and BP Neural Network

Hongxing Li[1], Wenjin Li[2], Jingrong Shu[3], Chen Ye-Hui[1*]
[1,2,3] Electronic Communication Engineering College
Anhui Xinhua University, Hefei
Anhui, 230088, China

ABSTRACT: *At present, the accuracy and robustness of unsupervised learning algorithm for multidimensional outlier detection are not good, this paper conducts a supervised learning for partial data by supervised BP neural network. Reliability of data is guaranteed by weighted entropy, then the author learns the relationship between non outliers and outliers through BP neural network, forecasts test data by trained BP neural network, finally compares the change of the data position before and after the prediction, take it as multidimensional data outlier distance, so as to complete this test. This article carries on the test experiment on the data set of multiple data sets. Compared with the traditional LOF and PSO methods, the experimental results show that the robustness and accuracy of the proposed algorithm for outlier detection of multidimensional data are better, which is suitable for all kinds of outlier detection of multidimensional data.*

Keywords: Weighted Entropy, BP Neural Network, Multidimensional Data, Outlier Detection

Received: 19 March 2017, Revised 1 May 2017, Accepted 8 May 2017

1. Introduction

With the rapid development of computer technology and intelligent technology, the data is growing in a massive way, the traditional data processing methods have been unable to meet the massive growth of data, so data mining technology arises at the historic moment. The data mining technology can dig out the meaningful intrinsic information from the massive original data through the intelligent mining algorithm, and carries on the decision-making through the intrinsic information for the multi-dimensional massive data [1]. In the data mining algorithm, outliers have great impact on the algorithm of intelligent mining, it usually reduces the adaptability and robustness of data mining algorithms, so the outlier detection and analysis of data mining is one of the most important tasks. Outlier detection aims to find abnormal isolated and sparse data points hidden in the data, the behavior of these abnormal data points are inconsistent with most of the data, which is abnormal, it will play a role of interference in data mining, so the outlier detection is one of the important steps in preprocessing of data mining [2]. Usually, due to

inconsistent patterns hidden in large data need strong background knowledge to be able to detect, it often loses important information by blind rejection, it is the focus of current research on how to use intelligent methods to detect outliers. So far, researchers have proposed a large number of intelligent algorithms for outlier detection. These methods mainly include five categories [3]:

(1) Statistical Methods. Statistical methods construct a statistical model by treating data processing, and then calculate the probability of each data object which conform to the statistical model. A data object that does not conform to the model is detected by hypothesis testing, and these objects are treated as outliers, finally subsequent processing is conducted.

(2) Distance Method. Construct the distance between feature data through the analysis of the background of data object, set a threshold for the distance between the object, when the distance is greater than the threshold, the data object is regarded as an outlier, and the subsequent processing is carried out.

(3) Depth Approach. For higher dimensional data objects, the depth value of each data is calculated by defining the depth of the data object, and the lower the depth value is, the higher the outliers are considered.

(4) Density Approach. Calculate the distance between data objects and give a distance range, calculate the distance within the range of the number of data objects, and thus calculate the data object density, the lower density region is, and the higher degree of outlier is. The most classical algorithm LOF of outlier detection achieves it by calculating the density of data.

(5) Method based on Swarm Intelligence. Swarm intelligence methods including particle swarm optimization, ant colony and genetic algorithm, the method defines a function of outlier degree according to the characteristics of outliers, and gradually optimizes the outlier degree function by randomization group, so as to detect the outliers.

In the present study algorithm of outliers, statistical method, distance method and depth method for multidimensional data are difficult to define and the complexity of computing time is high; Density method requires a large computational space and storage space to store intermediate results of density calculation; Swarm intelligence method requires higher data distribution, if the distribution of data objects is chaotic, it is easy to fall into local minimum, the initial conditions are harsh, the scope of application is not high [4]. In this paper, the data objects are divided into training set and test set by introducing weighted entropy, then the training set is trained by the BP neural network, so BP neural network can predict the value of data objects, and then determine whether it is outliers by comparing data object value before and after prediction. The method obtains the intrinsic relationship of multidimensional data by neural network, it can train different neural networks with different data, has a strong applicability, and it defines outlier judgment through the internal relation, false detection rate is lower.

2. Weighted Entropy and BP Neural Network

2.1 Weighted Entropy

Entropy is introduced into the information theory by American mathematician Shannon in the study of thermodynamics and entropy is defined as the random variable information [5]. The information entropy can be used to describe the degree of disorder of the information, and the outliers can be used as a state of the data. For a one-dimensional random variable X, if the symbol set is $\{a_1, a_1, a_n\}$ the number of random events is n, and the probability of occurrence of a_i is $p(a_i)$.The information entropy of the random variable can be defined as:

$$H(X) = -\sum_{i=1}^{n} p(a_i) \log p(a_i) \quad p(a_i) \geq 0, \sum_{i=1}^{n} p(a_i) = 1 \tag{1}$$

When the data object is a random variable, the random variable is composed of several discrete variables $X = (X_1, X_2, .., X_N)$, and each random variable $X_i (i = 1, 2, ..., N)$ is assumed to be independent, independent distribution probability can be satisfied:

$$p(X_1, X_2, ..., X_N) = p(X_1) \times p(X_2) \times ... \times p(X_N) \tag{2}$$

It is assumed that the set of symbols of different random variables can be represented by $\{a_1, a_2, ... a_n\}$, and then the

information entropy of multidimensional random variables can be defined as:

$$H(X) = H(X^N) = -\sum_{X^1}^{X^N} p(X)\log(X) = -\sum_{X^1}^{X^N} p(a_i)\log(a_i) \quad (3)$$

Among them, X^N represents the sum of all *N* dimensional random variables.

Multidimensional information entropy can effectively represent the disorder degree of a random variable, but the definition is defined according to the data, and each data variable is not connected. In the actual data set, the random events will be restricted by some conditions in the process. In order to introduce restrictions to more fully describe the multidimensional random variables, for each event, a nonnegative real number is defined to form weighted entropy for each event. The weighted entropy can be used to describe the degree of disorder of the random variables in the real environment, and it is more suitable for the outlier detection [6]. Based on the above description, the weighted entropy can be defined as:

$$H_\omega(X^N) = -\sum_{X^1}^{X^N} \omega_i p(a_i)\log(a_i) \quad (4)$$

Through the above conditions for the characterization of outliers, it is more statistical significance, can make the outlier detection more objective, better fitness. In the actual outlier detection, the definition of *A* and *B* should be defined by the definition of outliers to form the weighted entropy of the described data.

2.2 BP Neural Network

Neural network is a mathematical model to simulate the human brain neurons; the model can effectively solve the nonlinear problem, neurons connect synapses between neurons in the brain by weight, a large number of neurons and their weights can be used to form a highly complex nonlinear adaptive system, it can effectively complete tasks of control, decision-making, classification, regression and others through the system. For some problems that cannot use the appropriate formula and rules, we can use neural networks to perfectly solve, because it has strong flexibility and adaptability [7].

In the neural network, the most commonly used is the BP neural network, BP neural network carry on back-propagation through error, and constantly update the weights between neurons, so that the neural network gradually become reliable. For the outlier detection of multidimensional data, it is difficult to visualize the rule representation, so the neural network can be used to solve the problem. BP neural network is supervised learning, in outlier detection, we can first select the multidimensional data objects in non-outliers through experience, these points accounted for less than 10% of the data. Take these non-outliers as training data set for training the BP neural network; the remaining 90% data set as test data. Determine whether the data object is an outlier through the changes before and after BP neural network trained by test data, and combined with the weighted entropy [8].

According to the conventional neural network structure, this paper constructs a three layer neural network for outlier detection, which includes input layer, hidden layer and output layer. Each layer is composed of *N* neurons, which correspond to each dimension of multidimensional data object, and the number of neurons in the output layer is the same as the input layer, the hidden layer is made up of *M* neurons $M > N$, and the weights of the neurons are all connected. The following figure 1 gives the architecture of the three layer neural network of the outlier detection in this paper.

In the forward propagation of the neural network, the conduction is carried out by the following formula:

$$X_j = \sum_{i=1}^{N} X_i \omega_{ij} + b_{ij} \quad (5)$$

$$y_j = \frac{1}{1+e^{-X_j}} \quad (6)$$

Among them, ω_{ij} represents the weights between the input layer and the hidden layer, b_{ij} is additive bias, the sigmoid function

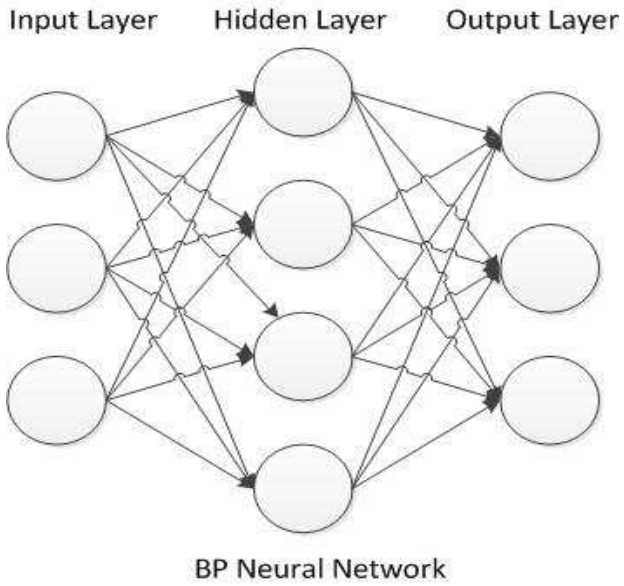


Figure 1. The architecture of the neural network of the outlier detection in three layers

is used between the hidden layer and the output layer, and the output y_j of the neural network is obtained through the output *Xi*.

BP back-propagation algorithm [9] is used to construct the mean square error between the actual value and the predicted value, and the mean square error is propagated back through the neural network. Update the weights of neural network by mean square error ω_{ij} and b_{ij}. Mean square error is defined as:

$$E=\frac{1}{N}\sum_{j=1}^{N}(y_j-d_j)^2 \tag{7}$$

By solving the mean square error, the relationship between the mean square error and the weight ω_{ij} and b_{ij} is established. Use the following formula to solve:

$$\frac{\partial E}{\partial y_j}=y_j-d_j \tag{8}$$

$$\frac{\partial E}{\partial \omega_{ji}}=\frac{\partial E}{\partial x_j}\frac{\partial x_j}{\partial \omega_{ji}}=\frac{\partial E}{\partial x_j}y_j \tag{9}$$

$$\frac{\partial E}{\partial x_j}=\frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial x_j} \tag{10}$$

$$\frac{\partial y_j}{\partial x_j}=y_j(1-y_j) \tag{11}$$

$$\frac{\partial E}{\partial \omega_{ji}}=y_j(1-y_j)(y_j-d_j)y_i \tag{12}$$

$$\Delta\omega_{ji} = -l\frac{\partial E}{\partial \omega_{ji}} \tag{13}$$

$$\omega_{ji} = \omega_{ji} + \Delta\omega_{ji} \tag{14}$$

Among them, l is the learning rate, the weight of the differential update weights through the learning rate and mean square error, after several iterations, the BP neural network can obtain optimal weights, this weight is suitable for testing the remaining 90% data sets, so as to find out the abnormal points.

For the test data set object, the actual data value is X^i, the prediction result is Y^i, you can define the extent of each data set object by the following:

$$O = \sum_{i=1}^{N} \frac{\left|X^i - Y^i\right|}{Y^i} H_\omega(X^i) \tag{15}$$

Finally, N dimensional data objects in the test data set can calculate outlier degree of each object through the algorithm; it can limit the outlier degree by the threshold, take the outlier data object as outliers, and completes the outlier detection.

3. Experiment and Result Analysis of Outlier Detection

3.1 Experimental Data Set

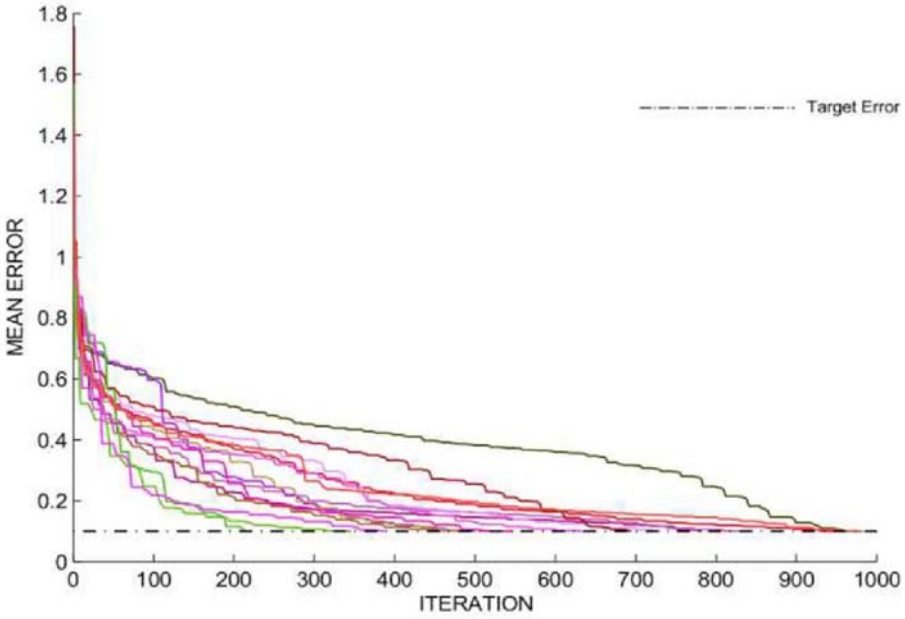
In order to verify the effectiveness of the algorithm proposed in this paper, the proposed BP neural network algorithm and the traditional clustering algorithms [10] such as, LOF [11], PSO [12], and GA [13] are compared on the UCI data set. In [11], the LOF algorithm using Hockey data, it has no objective criteria for outlier detection, in order to complete the comparison algorithm better, this paper extends the data set of KDDCUP99, data set for outlier detection in UCI [14], select the appropriate data to reconstruct the data set used in this paper.

The rules for reconstructing data are as follows: Select the data object in large group as a normal data object, select the data object in small group as an outlier, which can effectively expand the differences between different categories, so as to construct suitable for multidimensional outlier detection training data set and test data set, and it has a corresponding evaluation standard. Because the UCI data set contains a large number of network connection records, each record contains the attributes of more than 41, according to the experience of test results, 15 of these attributes have a greater impact on the results, so this paper chooses the 15 great impact properties, and select 10050 data sets from the data set. The data set contains a total of 550 outliers in the data set, choose 10% as the training data, the remaining 90% as test data, ensure that training data have no outliers. The data sets are used to test the effectiveness and efficiency of multidimensional outlier detection.

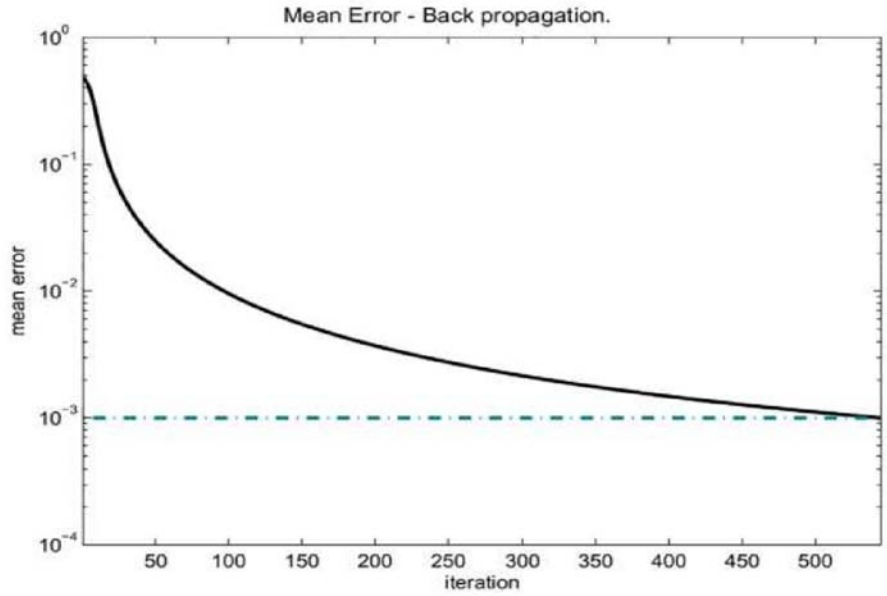
3.2 Experimental Results

In the course of the experiment, 15 meaningful attributes are selected in this paper, and the dimension of detection is 15, so the input and output of the neural network has a total of 15 neurons. In general, the number of neurons in the hidden layer needs to be greater than or equal to the number of the input neurons, in this paper, the number of neurons selected is 20. In summary, this paper constructs a 15-20-15 BP neural network for multidimensional data outlier detection, and BP algorithm is used to optimize the parameters of the BP neural network. In the initialization process, BP neural network is evaluated by random values. Using sigmoid function given by formula (6) for the output of the incentive, using mean square error given by formula (7) as the error propagation neural network, the maximum number of iterations is 1000 times. The following figure 2 (1) gives the results of the variation of the error of the neural network training with the 15 different attributes, 2 (2) gives the overall decline process with the BP algorithm of mean square error.

After iterative update of BP neural network, the neural network can effectively complete the forecast data in each attribute, the remaining 90% data is used as the test data set, and predicted data attribute value is obtained after neural network predicts data. The weighted entropy is calculated by formula (4) and (15), and the degree of outlier is calculated. according to the predicted



(1) Iterative error results of 15 different attributes



(2) The iterative results of the mean square error

Figure 2. Test results of BP neural network in training data set

attribute value and the actual attribute value. Figure 3 shows the comparison of the number of outliers and the number of outliers calculated by neural network that trained in different times.

In order to give a more intuitive view of the advantages of the neural network in this paper, the author compares it with the common clustering algorithm, LOF, SPO, GA and so on, the following table gives the comparison of the accuracy and time efficiency of outlier detection.

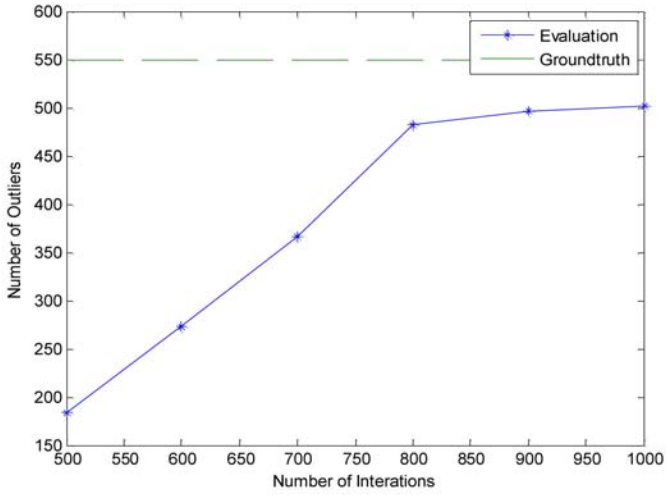


Figure 3. Comparison of the number of outliers predicted by weighted entropy of neural network and the actual number

Algorithm	**Correct rate of outlier detection**	**Detection efficiency of outliers**
Algorithm in this paper	91.27%	136720ms
K-means	67.38%	2938ms
LOF	75.28%	5628ms
SPO	83.27%	2673ms
GA	81.38%	2698ms

Table 1. Comparison of accuracy and time efficiency of different algorithms in outlier detection

3.3 Results Analysis

The experimental results show that, through the combination of data sets to test the validity of this method, test 15 dimensional attribute data in BP neural network. In the training of the BP neural networks, the mean square error and the mean square error of all attributes are tested. The test results show that the BP neural network in this paper can effectively detect the outliers of multi dimension data under limited training and error back propagation. In comparison with the actual data, along with the increase of the number of training, the effect of neural network algorithm proposed for outlier detection has also been significantly improved, when the number reaches a certain number, the detection of outliers gradually slowed down, the algorithm in this paper, falls into the bottleneck of overfitting. In fact, compared with other algorithms it can be seen that this algorithm can detect the outliers in the maximum limit, and because the algorithm have training and testing process, so it's time is more than traditional method of unsupervised learning. But compared with the accuracy of outlier detection, the correct rate of the algorithm has been significantly improved. By sacrificing the time complexity, the correct rate of outliers is significantly improved, which is of great significance in outlier detection of multidimensional data.

4. Conclusions

With the rapid development of data mining technology, one of the most important research directions in data mining is to find out outliers from the data to be mined. Outliers will have a greater impact on the various data mining algorithms, so it is very important to detect the outliers from the data set before data mining. Starting from supervised learning, this article trains an

effective BP neural network by extracting a small number of non-outliers in the data, then calculates the weighted entropy through predictive value of neural network and actual result value, calculates the degree of outlier of each data object, determine whether the data object belongs to outliers through the threshold. Compared with other traditional algorithms, the algorithm can significantly improve the accuracy of outlier detection on the basis of sacrificing a small amount of training time. In the actual use, with the increase of data size, the amount of data used for training is far less than the amount of test data, so that the time complexity of the algorithm tends to be balanced with other traditional algorithms. In conclusion, the proposed algorithm can effectively solve the massive multidimensional data outlier detection and data mining, which has important practical significance.

Acknowledgements

We acknowledge the support of National Natural Science Foundation of China. research on nonlinear dynamic characteristics at large angle of attack for free-flying object with asymmetric fin and its application. 11272356.

References

[1] Wang, J. S., Chiang, J. C. (2008)A cluster validity measure with outlier detection for support vector clustering. *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society*, 38 (1) 78-89.

[2] Zhao, W., Chen, D., Hu, S. (2004). Detection of outlier and a robust BP algorithm against outlier. *Computers & Chemical Engineering*, 28 (8) 1403-1408.

[3] Li, N., Li, P., Shi, X., et al. (2010). Outlier identify based on BP neural network in dam safety monitoring, International Asia Conference on Informatics in Control, *Automation and Robotics*. IEEE, 210-214.

[4] Ker, A D., Pevný, T. (2014).The Steganographer is the Outlier: Realistic Large-Scale Steganalysis, *IEEE Transactions on Information Forensics & Security*, 9 (9) 1424-1435.

[5] Dong, Y., Ning, H. (2012). A Modal for Outlier Removal by MLBP Neural Network Based on Adaptive Performance Function, *Communications and Information Processing*. 663-671.

[6] Kong, D. Penalized Regression Methods with Application to Domain Selection and Outlier Detection.

[7] Achour, S., Rinard, M C. (2015). Approximate computation with outlier detection in Topaz, *Acm Sigplan Notices*, 50 (10) 711-730.

[8] Paguio, H J S. (2013). Artificial neural network application for magnetic core width prediction and modeling for magnetic disk drive manufacture.

[9] Guang-Qiang, L I., Liu, Q L., Deng, M. (2009). A BP Neural Networks Based Spatial Outliers Detecting Method. *Journal of Geomatics Science & Technology*.

[10] Li, N., Li, P., Shi, X., et al. (2010). *Outlier identify based on BP neural network in dam safety monitoring*, 2. 210-214.

[11] Xiao-Yan, L U., Jing-Chao, L U., Zhou L R., et al. (2010). BP Neural Network Adaptive Kalman Filtering in Existence of Outliers, *Computer Simulation.*

[12] Jian, L I., Yan, B P., Jun, L i. (2008). Memory-effect-based Local Outlier Detection Algorithm, *Computer Engineering*, 34 (12) 4-6.

[13] Tseng, F M., Yu, H. C., Tzeng, G H. (2002). Combining neural network model with seasonal time series ARIMA model, *Technological Forecasting & Social Change*, 69 (1) 71-87.

[14] Li ,Y., Gao, Q., Bøegh, J. (2013). The Evaluation of Service Trustworthiness Based on BP-Neural Network, Trustworthy Computing and Services. Springer Berlin Heidelberg, 185-190.

Author Bibliography

Hongxing Li, Female, 1982.11, 15155189308, 46713779@qq.com, 555, Wangjiang West Road, Anhui, Hefei, Anhui Xinhua University, Master of Engineering, Lecturer

Wenjin Li, Female, 1982.01, 138661542940liwenjin17@163.com, 555, Wangjiang West Road, Anhui, Hefei, Anhui Xinhua University, Master of Engineering, Lecturer