TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.5, May 2014, pp. 4091 ~ 4100
DOI: http://dx.doi.org/10.11591/telkomnika.v12i5.4388

Based on Weighted Gauss-Newton Neural Network Algorithm for Uneven Forestry Information Text Classification

Yu Chen*, Liwei Xu
Department of information and computer engineering, Northeast forestry university, China
Harbin city of Heilongjiang Province, 26 Hexing Road,
Xiangfang District, Northeast Forestry University, 150040
*Corresponding author, email: xuliwei475273608@163.com

Abstract

In order to deal with the problem of low categorization accuracy of minority class of the uneven forestry information text classification algorithm, this paper puts forward the uneven forestry information text classification algorithm based on weighted Gauss-Newton neural network, on the basis of weighted Gauss-Newton algorithm, the algorithm is proved via singular value decomposition principle. The experimental result shows that the algorithm has higher classification accuracy of majority class and minority class than algorithm of common classification. The algorithm expands a new method for the research on the uneven forestry information text classification algorithm.

Keywords: *text classification, weighted Gauss-Newton, iterative algorithm*

1. Introduction

Forestry information resources are very rich in China, with more and more people use the Internet, forestry information obtained also showed an upward trend, but in real life, people just want the Internet to provide a small part of forestry information, therefore the forestry information text classification technology emerge as the time requires.

This paper puts forward the uneven forestry information text classification algorithm based on weighted Gauss-Newton neural network. Firstly, pretreatment of uneven forestry information text using ICTCLAS Chinese word segmentation system of the Chinese Academy of Sciences (segmentation and to stop words), secondly, using the classical TF-IDF formula to calculate the eigenvalues of text word ,constitute the initial text feature matrix, Then, by the method of principal component analysis to reduce the dimensionality of the feature matrix；Finally, the dimensionality reduction characteristic matrix for training, construct weighted Gauss-Newton neural network classifier, in order to achieve the purpose of classification. Through a lot of experiments demonstrate that the algorithm has reached the expected goal, the classification of the minority class and majority class has higher correct rate. Because of the specificity of the uneven text [1-3], through the global accuracy or error rate to evaluate the performance of classifier are not enough, therefore, the geometric mean formula are introduced to consider the classification performance of the minority class and majority class samples [4]. The classification performance of the algorithm is significantly higher than that of the classical classification method of uneven forestry information text, it provides a new method for uneven forestry information text classification.

2. Key Technology of the Uneven Forestry Information Text Classification Algorithm Based on Weighted Gauss-Newton Neural Network

2.1. Representation of uneven forestry information text

Pretreatment of uneven forestry information text using ICTCLAS Chinese word segmentation system of the Chinese Academy of Sciences (segmentation and to stop words), count the weight of word of uneven forestry information text, constitute the initial text feature matrix.

Received October 15, 2013; Revised December 26, 2013; Accepted January 19, 2014

The assumption that the total number of all characteristics of uneven forestry information text is n, formation the n-dimensional vector space, Each uneven forestry information text d is represented as a n-dimensional feature vector.

$$V\text{（d）} = (T_1, W_1(d); T_2, W_2(d); \cdots\cdots; T_n, W_n(d)) \quad (1)$$

Here, T_i is the text segmentation of uneven forestry information, $W_i(d)$ is the weight of T_i in text d, using the TF-IDF formula to calculate the weight of text word [5].

$$\mathrm{w}_{\mathrm{i}}(d) = \frac{TF(t_i) \times IDF(t_i)}{\sqrt{\sum_{i=1}^{n}(TF(t_i) \times IDF(t_i))^2}} = \frac{TF(\mathrm{t}_i) \times \log(\frac{N}{n_i} + L)}{\sqrt{\sum_{i=1}^{n}(TF(t_i) \times \log(\frac{N}{n_i} + L))^2}} \quad (2)$$

In the formula (2), $\mathrm{w}_i(d)$ represents the weight of feature words T_i, $TF(\mathrm{t}_i)$ is the number of feature words T_i that appears in the text d, N represents the total number of uneven forestry information text, n_i is the number of feature words T_i appeared in the uneven forestry information text.

2.2. Uneven Forestry Information Text Feature Selection

The text feature matrix dimension reduction, the selection of principal component analysis.

Suppose there are n samples of text, each sample has p eigenvalues $X_1, X_2, ...X_P$, get the original data feature matrix [6].

$$X = \begin{bmatrix} x_{11} & x_{12} & & x_{1p} \\ x_{21} & x_{22} & & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & & x_{np} \end{bmatrix} = (X_1 \quad X_2 \quad \cdots \quad X_p) \; X_{\mathrm{i}} = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ \\ x_{ni} \end{bmatrix} \quad , \; i = 1,2,\cdots, \; \mathrm{p} \quad (3)$$

Text feature matrix X is a linear combination of p vectors $X_1, X_2, ...X_P$, Principal component scores generated.

$$\begin{cases} F_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p \\ F_2 = a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p \\ \vdots \\ F_p = a_{1p}X_1 + a_{2p}X_2 + \cdots + a_{pp}X_p \end{cases} \quad (4)$$

Equals to:

$$F_{\mathrm{i}} = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{pi}X_p \quad , \quad i = 1,2,\cdots p \quad (5)$$

Constraint for coefficient of:

$$a_i = (a_{1i}, \; a_{2i}, \cdots, \; a_{pi})$$
$$a_{1i}^2 + a_{2i}^2 + \cdots + a_{ni}^2 = 1, \quad i = 1,2,,\cdots, \; p \quad (6)$$

Use the following formula to calculate the covariance matrix of text feature matrix $S=(s_{ij})_{p\times p}$

Among,

$$s_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j) \quad , \quad i,j=1,2,\cdots, \quad p \tag{7}$$

Calculating the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ of covariance matrix S and the corresponding eigenvector.

$$a_1=\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix}, \quad a_2=\begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p2} \end{bmatrix}, \cdots, \quad a_p=\begin{bmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{pp} \end{bmatrix} \tag{8}$$

The i -th principle component scores of feature matrix is $F_i = a'_i X \quad , \quad i=1,2,\cdots, \quad p$

By calculating the contribution rate and the cumulative contribution rate to determine all of the main components which should be selected for experimental evaluation.

$$\alpha_i=\frac{\lambda_i}{\sum_{j=1}^{p}\lambda_i} \quad \text{and} \quad G(r)=\frac{\sum_{j=1}^{r}\lambda_i}{\sum_{j=1}^{p}\lambda_i} \tag{9}$$

In the experiment, the extraction of the cumulative contribution rate of 99% of the main component, calculation of n samples in the selected r forestry information principal components scor.

$$F_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{pi}X_p \quad , \quad i=1,2,\cdots, \quad r \tag{10}$$

2.3. Weighted Gauss-Newton Algorithm

Commonly used classification methods of uneven forestry information text are support vector machines, Bayesian, decision tree, etc. These classic uneven forestry information text classification algorithm for the minority class classification accuracy rate is very low, the experimental results of uneven forestry information text classification algorithm based on weighted Gauss-Newton neural network have been better, Improve the accuracy of the minority class classification [7, 8].

Newton's method using the main idea of the second-order Taylor expansion of the objective function, and then find its minimization [9].

Assuming that $f(x)$ twice differentiable, $x_k \in R^n$, $\nabla^2 f(x_k)$ is positive definite Hess matrix, using the Taylor expansion of $f(x)$, formula for the type [10].

$$f(x_k+s) \approx f(x_k)+\nabla f(x_k)^T s+\frac{1}{2}s^T\nabla^2 f(x_k)s \tag{11}$$

The above formula, $s=x-x_k$, the minimum value can be obtained by the following formula.

$$x_{k+1}=x_k-T_k^{-1}t_k \tag{12}$$

The above formula is Newton iteration formula, in the formula $T_k = \nabla^2 f(x_k)$, $t_k = \nabla f(x_k)$, that is to say T_k represents the second derivative and t_k represents first derivative of the function [11].

Newton method, the selection of initial point is very important, If the initial point away from the optimum value obtained is far from the last, the second derivative matrix is not necessarily positive definite, therefore, the search direction is not necessarily decline, and the final result will not accurate enough [12, 13].

Gauss-Newton is the iterative algorithm for unconstrained minimization, the basic idea is the objective function of the least squares problem ignore second-order information items $S(x)$, the two model of objective function turns into:

$$\overline{m_k}(x) = \frac{1}{2} r(x_k)^T r(x_k) + \left(J(x_k)^T r(x_k)\right)^T (x - x_k) + \frac{1}{2}(x - x_k)^T (J(x_k)^T J(x_k))(x - x_k) \qquad (13)$$

With $\mathrm{v}^T(x)$ representing the output value of the error, the error function is expressed as:

$$\mathrm{f(x)} = \sum_{i=1}^{n} v_i^{\,2}(x) = v^T(x) v(x) \qquad (14)$$

Gradient of $f(x)$ is:

$$\nabla f(x) = 2J^T(x) v(x) \qquad (15)$$

Hess matrix of $f(x)$ is:

$$\nabla^2 f(x) = 2J^T(x) J(x) + 2S(x) \qquad (16)$$

In the two formulas, $J(x)$ is Jacobi matrix, put the last two equations into formula 14 to obtain iteration formula of Gauss – Newton.

$$x_{k+1} = x_k - \left[J^T(x_k) J(x_k)\right]^{-1} J^T(x_k) v(x_k) \qquad (17)$$

Compare Gauss-Newton with Newton method, there’s no need to calculate $\nabla^2 f(x)$, avoid the second order matrix is not positive definite and search direction does not necessarily decline, but in the formula, $J^T(x_k) J^T(x_k)$ still irreversible, So the algorithm may not converge. The classification of directly using M algorithm is ineffective, to solve the above problems, the Gaussian iteration formula, $J(x)^T J(x)$ adds a parameterized unit array λI to make the Gauss - Newton algorithm has better regularization nature, overcome when S is singular, the algorithm converges to a case of non-resident. Formula as follows.

$$x_{k+1} = x_k - \left[J^T(x_k) J(x_k) + \lambda I\right]^{-1} J^T(x_k) v(x_k) \qquad (18)$$

The above formula, I is a $n \times n$ unit matrix, $\lambda > 0$, λ is the regularization parameter. To make Gauss - Newton algorithm with global convergence, then add a one-dimensional search factor α with damping function to obtain the formula.

$$x_{k+1} = x_k - \alpha_k \left[J^T(x_k) J(x_k) + \lambda I\right]^{-1} J^T(x_k) v(x_k) \qquad (19)$$

In the above formula, α_k I is a one-dimensional search factor, expressed as follows.

$$\alpha_k = \frac{\left\| J(x_k)^T v(x_k) \right\|^2}{\left\| J(x_k) J^T(x_k) v(x_k) \right\|^2} \tag{20}$$

Parameter λ_k is determined by the selection as follows:

$$\lambda_k = \alpha_k (\theta \| v(x_k) \| + (1-\theta) \| J^T(x_k) v(x_k) \|, \quad \theta \in (0,1) \tag{21}$$

Here it must be noted that the selection of parameter θ should be larger, because of the very large dimension of text feature matrix, So $\left\| J(x_k)^T v(x_k) \right\|$ will have a significant number of norm, and the parameter $\left\| v(x_k) \right\|$ norm of value will be smaller, so must ensure that $\theta > 1-\theta$.

Although the different parameters and adjustment of $J(x)^T J(x)$, the Gauss - Newton algorithm has better convergence, there are still some restriction factors makes the classification effect poor, so the type and then weighted processing, join the weight matrix ω_k to above formula, reduce the error feature matrix dimensionality reduction on the classification of the impact, the classification performance is improved, Weighted Gauss - Newton iterative formula is as follows:

$$x_{k+1} = x_k - \omega_k \alpha_k \left[J^T(x_k) J(x_k) + \lambda I \right]^{-1} J^T(x_k) v(x_k) \tag{22}$$

In the above formula, Weight matrix is as follows:

$$\omega = \beta.diag(\left| \frac{1}{\Delta\rho_1^1} \right|, \left| \frac{1}{\Delta\rho_{2^1}^1} \right|, \ldots, \left| \frac{1}{\Delta\rho_N^1} \right|) \tag{23}$$

$\Delta\rho_i^1$ is the i-th component of calculated $-\left[J^T(x_k) J(x_k) + \lambda I \right]^{-1} J^T(x_k) v(x_k)$, β is the scale factor, formula is as follows:

$$\beta^2 . \sum_{i=1}^{N} \frac{1}{\left| \Delta\rho_i^1 \right|^2} = N^2 \tag{24}$$

Formula 24 as the iterative method of weighted Gauss-Newton algorithm.

The following stability of weighted Gauss-Newton iterative method is proved. The iterative formula 24 correspond to a linear least squares problem of equations.

$$s = -\alpha\omega [J^T(x_k) J(x_k) + \lambda I] J^T(x_k) v(x_k) \tag{25}$$

The iterative formula 19 correspond to a linear least squares problem for the equations.

$$s_N = -[J^T(x_k) J(x_k)]^{-1} J^T(x) v(x_k) \tag{26}$$

In the formula 27 meets $\alpha = \alpha_k$.

2.4. Text Classifier Performance Evaluation of Uneven Forestry Information Text

Use only the global accuracy or error to evaluate the imbalanced data classifier is one-sided, therefore, the introduction of the following formula, considering the classification performance of minority and majority class [14].

The correct rate of minority class samples, TP said the minority class is the number of correctly classified, FN refers to the number of misclassification of minority class to the majority class.

$$\text{Sensitivity}=\frac{TP}{(TP+FN)} \tag{27}$$

The correct rate of majority class samples, TN said the majority class is the number of correctly classified, FP refers to the number of misclassification of majority class to the minority class.

$$Specificit\ y=\frac{TN}{(FP+TN)} \tag{28}$$

$\Pr ecision$ represents for minority class precision.

$$\Pr ecision=\frac{TP}{(FP+TP)} \tag{29}$$

Correct rate of geometric mean $G-\text{mean}$.

$$G=\sqrt{Sensitivity\cdot Specificity} \tag{30}$$

Minority class's $F-measure$.

$$F=\frac{2\times Sensitivit\ y\times \Pr ecision}{Sensitivit\ y+\Pr ecision} \tag{31}$$

3. Experimental Results

The following table shows the selected experimental samples.

Table 1. Selection Table of Uneven Forestry Information Text

	flowers	trees	insects	Soil type	water class
the number of training sample	1000	50	1000	1000	50
the number of test sample	50	50	50	50	50

As shown in Table 1, select five categories of uneven forestry information, flowers, trees, insects, soil type, water class, technical point uneven data refers to the different classes show unequal distribution of sample sets, so choose flowers, insects, soil three kinds of samples as the majority class, the election of 1000 samples. Class trees, water samples were minority class types, each of 50 samples, each select 50 samples to test.

Preliminary laboratory, algorithm design process, the Gauss-Newton algorithm add three parameters in order, λI， α， w , The formation of iterative formula of weighted Gauss-Newton algorithm, algorithm improvement process, results show below.

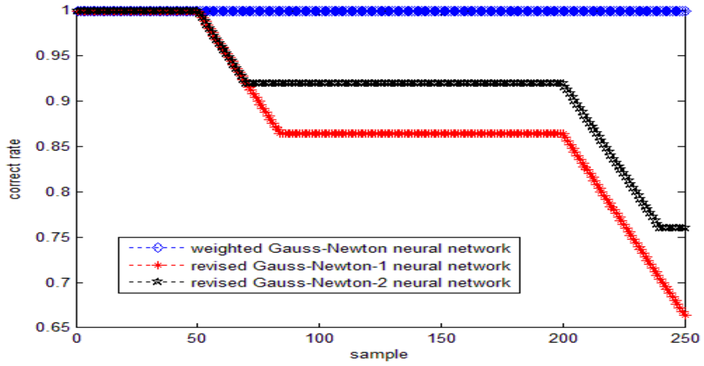

Figure 1. Weighted Gauss-Newton Neural Network Comparison Chart

Revised Gauss-Newton-1 stands for algorithm add the first parameter λI , using formula 20 to iterate. Revised Gauss-Newton-2 represents the first time on the basis of parameters have been added, add a second parameter α, using formula 21 to iterate. Weighted Gauss-Newton algorithm represents the third parameter weight matrix w be added, using formula 24 to iterate, the final results, with three parameters gradually add into the algorithm, the rate of correct classification improve gradually, with second parameters, the correct rate increased marginally, therefore, adding third parameters, the weighted Gauss-Newton algorithm with third parameters to improve the accuracy of uneven of forestry information classification, during the experiment, compare weighted Gauss-Newton neural network algorithm with commonly used classification algorithm of uneven forestry information text.

Uneven initial training sample dimension of feature matrix is 3100×1127, the initial test sample dimension of feature matrix is 250×1127, these two matrices form a new dimensionality reduction dimension 3100×213 and 250×213 of feature matrix. Using four methods of weighted Gauss-Newton neural networks, decision trees, Bayesian, support vector machine classification, select the same training and test samples, test results are shown below, abscissa is test samples category of uneven forestry information text, the vertical axis is the correct rate of each type of sample classification.1 represents for type of flowers, 2 represents for type of trees, 3 represents for type of insects, 4 and 5 represent for type of soil and water.

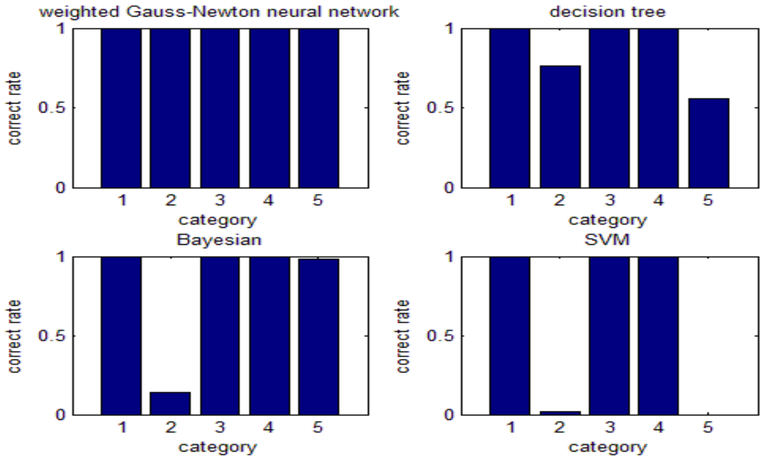

Figure 2. Four Classification Algorithms for Uneven Forestry Information Text Classification Accuracy Schematic

In Figure 2, results of uneven forestry information text classifiers, in decision trees and Bayesian classification algorithm the minority class sample classification accuracy rate is low, Bayesian for the tree sample recognition rate is very low . Support vector machine classifier classification accuracy for minority class is almost zero. Weighted Gauss-Newton neural network classification accuracy for minority class can reach 100%, and for the majority of the class classification accuracy is also high.

The above chart reflect different aspects of classifier performance, in order to measure classification performance of classifier more comprehensive, highlighting the importance of minority class in the classification process, using the comprehensive index judgments : F-measure and G-mean. Data will be divided into two categories, the majority class samples and the minority class samples.

In the following table, X1 represents majority class in reality and judgment is majority class, X2 represents majority class in reality but judgment is minority class, X3 represents minority class in reality and judgment is minority class, X4 represents minority class in reality but judgment is majority class.

Table 2. The Mixed Matrix of Test Sample Set of Four Classification Algorithm

	Weighted G-N	Decision tree	Bayesian	SVM
X1	150	150	150	150
X2	0	0	0	0
X3	100	66	56	1
X4	0	34	44	99

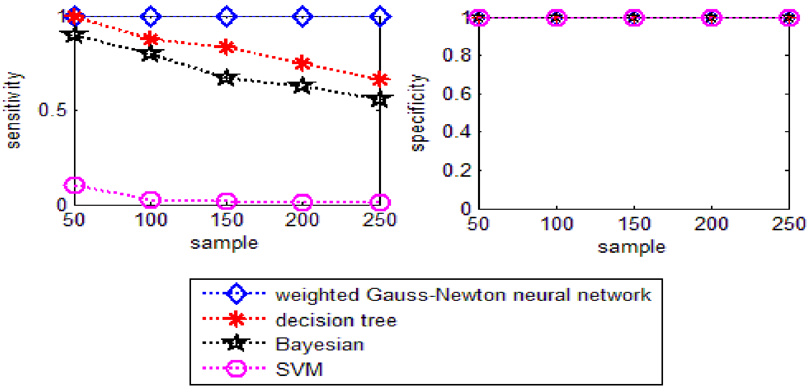

Figure 3. Change Figure of the Majority Class and Minority Class Accuracy

Figure 3 shows, as the sample increases, the majority class and minority class trend changes of correct rate, the weighted Gauss-Newton neural network with minority increase in accuracy of the sample does not change and always maintain at 100%. Decision tree, Bayesian, SVM as the sample increases, the correct classification rate of minority class show a decreasing trend, majority class classification results of four classifiers are better.

Table 3. Comprehensive Classification Effect of Four Kinds of Classifiers

	Precision	G-mean	F-measure
weighted G-N	1	1	1
decision tree	1	0.81	0.795
Bayesian	1	0.75	0.72
SVM	1	0.1	0.02

Table 3, G index consider the classification performance of the minority class and majority class samples, the value of G along with Sensitivity and Specificity value in [0 1] monotonically increasing, ensure that the two can make a large amount of both G value, weighted Gauss-Newton neural network classification performance is good. F-measure pays more attention to reflect the classification effect of the minority class. In summary, the weighted Gauss-Newton neural network classification performance is perfect, majority class and minority class classification accuracy equalization, decision tree, Bayesian and support vector machine classification performance of the minority class is poor, so comprehensive measure F-measure is small.

It's known from the experiment results, the proposed weighted Gauss-Newton neural network algorithm for five kinds of uneven forestry information text classification is precise and fast, especially for minority class sample classification accuracy is significantly higher than the commonly used classification algorithms, the algorithm correct classification rate is evenly distributed, classification ability is strong.

4. Summary

Based on the weighted Gauss-Newton neural network for uneven forestry information text classification algorithm, using the classical TF-IDF formula to calculate the eigenvalues of text words, constitute the initial text feature matrix, by the method of principal component analysis to reduce the dimensionality of the feature matrix, the formation of new uneven forestry information text feature matrix, in response to the feature of text. Through experiments show that the weighted Gauss-Newton neural network for uneven forestry information text classification algorithm, minority class classification accuracy is significantly higher than the classical method of decision tree classification, Bayesian, support vector machines, this algorithm provides a new algorithm for the study of uneven forestry information text classification, has high practical value.

Acknowledgements

Fundamental Research Funds for the Central Universities (DL12CB02); National 948 Project (2011-4-04); Heilongjiang Provincial Department of Education Science and Technology Research Project (12513016); Postdoctoral Fund of Heilongjiang Province. The Heilongjiang Province Natural Science Fund Project (F201347); Harbin technological innovation talent special fund project (2013RFQXJ100)

References

[1] Xie Na-na, Fang Bin, Wu Lei. Study of text categorization on imbalanced data. *Computer Engineering and Applications*. 2012; 6(1): 1-4.

[2] Duan Li-guo, Dip eng, Li Ai-Ping. A New Naïve Bayes Text Classification Algorithm. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2014; 12(2): 947-952.

[3] Pei-ying ZHANG. A HowNet-based Semantic Relatedness Kernel for Text Classification. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(4): 1909-1915.

[4] Fengfeng Bai. Uneven data sets based text classification technology research. *Software Development and Design*. 2009; 12(10): 21-29.

[5] Jiangli Duan. Research of Feature Selection and Weighting Algorithm in Text Classification System Based on SVM. Taiyuan: Taiyuan University of technology. 2011: 10-15.

[6] Yafei Wang. Text Categorization Method of Reducing Features. Changchun: Changchun University, 2010: 20-25.

[7] Zulin Hua, Wei Qian, Li Gu. Application of improved LM-BP neural network in water quality evaluation. *Water resources protection*. 2008; 24(4): 23-30.

[8] Deyun Chen, Yu Chen, Lili Wang, Xiaoyang Yu. A novel Gausss-Newton Image Reconstruction Algorithm for Electrical Capacitance Tomography System. *Chinese journal of electronics*. 2009; 37(4): 739-743.

[9] Subramanian PK, Xiu NH. Convergence Analysis of Gauss-Newton Methods for the Complementarity Problem. *Journal of Optimization Theory and Applications*. 1997; 94: 727-738.

[10] Yu Chen. Research on Inverse Problems Solving and Image Reconstruction Algorithm For Electrical Capacitance Tomography System. Harbin: Harbin University of Science and Technology. 2010: 57-60.

[11] Xiulan Chen, Jun Wei. Improved convergence analysis of Gauss Newton algorithm step. *Journal of Chinese science and technology innovation*. 2012; 1(1): 110-111.

[12] Subramanian PK. Gauss-Newton Methods for the Complementarity Problem. *Journal of Optimization Theory and Applications*. 1993; 77: 467-482.

[13] Rubanov NS. The layer-wise method and the back propagation hybrid approach to learning a feed forward neural network*. IEEE Trans. Nerual Networks*. 2000; 1(2): 295-305.

[14] Xinmin Tao, Furong Liu, Baoxiang Liu. Uneven data SVM classification algorithm and its application. Harbin: Heilongjiang Science and Technology Press. 2011: 14: 16.