Open Access

Feature Extraction Method of Ultrasonic Signal Based on Wavelet Coefficients Clustering

Feng Zhihong*[,1] Miao Changyun[2] and Bai Hua[2]

[1]*Engineering Practice Teaching Training Center, Tianjin Polytechnic University, Tianjin, P.R. China*

[2]*School of Electronics and Information Engineering, Tianjin Polytechnic University, Tianjin, P.R. China*

Abstract: Support Vector Machine can well solve the classification problem of small sample, but when the dimension of input feature vector is very large, the structure of classifier is complex, the training time is long, and the performance is decreased. To solve this problem, a feature extraction method based on wavelet coefficients clustering was proposed. All the wavelet coefficients were clustered, the energy value of wavelet coefficients in each clustering was calculated and used as the input feature vector of a classifier. The dimension of input data was greatly reduced and information of specific problem was reserved. Support Vector Machine was used to identify the defects in steel plate, experiment results showed that the proposed method has higher classification accuracy.

Keywords: Feature vector, Feature extraction, Support vector machine, Wavelet transform.

1. INTRODUCTION

In industrial field, ultrasonic nondestructive testing of workpieces is often required [1]. In the process of detection, a large number of ultrasonic echo signals need to be collected, and then processed and classified. In the process of classification recognition, the accuracy of classification depends mainly on extracted features and classifier used. The time-frequency localization of wavelet transform [2] makes it a common feature analysis method. However, the dimension of the discrete wavelet coefficients matrix of actual sampled signals is very large. To reduce the dimensionality, it is necessary to extract important features from the wavelet coefficients matrix. In the past, feature extraction methods based on wavelet transform often used wavelet coefficients of one scale or some scales as features [3-4]. But the wavelet coefficients at different scales obtained by wavelet transform represent the magnitude of signal components in different frequency ranges. Once the wavelet coefficients of some scales are discarded, the signal information contained in these scales will be lost and the accuracy of classification will be reduced. In addition, the position of wavelet coefficients is variable.

Based on above problems, a feature extraction method based on wavelet coefficients clustering was proposed in this paper. Firstly, fast wavelet transform of collected signals was used to obtain a wavelet coefficients matrix [5]. Secondly, in each row of the matrix, the coefficients containing larger signal information were extracted, and the coefficients with smaller information were merged, that is, clustering of wavelet coefficients at each scale, and finally, the wavelet coefficients energy value of each cluster was calculated and used as the input vector of a classifier. Proposed feature extraction method was combined with support vector machine classifier [6-7] to identify internal defects in medium heavy steel plates, the experiment results showed that proposed method has higher classification recognition accuracy.

2. FAST WAVELET TRANSFORM

The Daubechies wavelet basis has been widely used in signal analysis because it has the properties of ideal approximation accuracy and numerical stability [8]. Parameters corresponding to discrete wavelet decomposition are given by real sequence $(h_0, h_1, \cdots, h_N)$, and these values satisfy the following equation:

$$\sum_{n=0}^{N} h_n = \sqrt{2} \tag{1}$$

$$\sum_{n=0}^{N} (-1)^n n^k h_n = 0\,, k = 0, 1, \cdots, (N-1)/2 \tag{2}$$

$$\sum_{n=0}^{N-2k} h_n h_{n+2k} = 0,\ \ k = 1, 2, \cdots, (N-1)/2 \tag{3}$$

Therein, N is a singular natural number, the number of solutions is $2^{(N-1)/2}$.

It is assumed that the sampled signal obtained by ultrasonic sensor is a finite series of data $(a_0 = (a_{0n})(n = 0,1, \cdots, L-1))$ with a length of L , N is order of the Daubechies wavelet basis, and then the wavelet basis expansion has the following relation.

$$a_{(m+1)n} = \sum_{k=2n}^{2n+N} h_{k-2n} a_{mk} \tag{4}$$

Therein, $\mathrm{n} = \frac{1-\mathrm{N}}{2}, \frac{1-\mathrm{N}}{2} + 1, \cdots, \frac{\mathrm{L}}{2} - 1$ (L is even);

$$\mathrm{n} = \frac{1-\mathrm{N}}{2}, \frac{1-\mathrm{N}}{2} + 1, \cdots, \frac{\mathrm{L}-1}{2} \ (L \text{ is odd});$$

$$d_{(m+1)n} = \sum_{k=2n+1-N}^{2n+1} (-1)^k h_{2n+1-k} a_{mk} \tag{5}$$

*Address correspondence to this author at the Engineering Practice Teaching Training Center, Tianjin Polytechnic University, Tianjin 300387, P.R. China; Tel: +8613102066386; Email: fengzh@163.com

1874-4443/15 2015 Bentham Open

Therein, $\mathrm{n}=0,1,\cdots,\frac{\mathrm{L+N-3}}{2}$ (L is even); $\mathrm{n}=0,1,\cdots,\frac{\mathrm{L+N-2}}{2}$ (L is odd). They are the decomposition of information contained in a_0, and by computing equation (1), (2) and (3) we can obtain the following equation.

$$\boldsymbol{a_{0n}=\sum_{k=-\infty}^{\infty}[h_{n-2k}a_{1k}+(-1)^n h_{2k+1-n}d_{1k}],n\in Z} \quad (6)$$

Namely sequence a_0 can be reconstructed precisely by a_1 and d_1. The energy of sequence a_0 is divided into a_1 and d_1 and the following definition is given.

Definition 1: The energy $\|\mathrm{p}\|^2$ of a limited sequence $\mathrm{p}=(\mathrm{p_n})$ is defined as

$$\|\boldsymbol{p}\|^2=\boldsymbol{\sum_{n=-\infty}^{\infty}p_n^2} \quad (7)$$

If p is regarded as a vector, the square root of the energy $\|p\|^2$ is the Euclidean norm $\|p\|_2$ of p.

Theorem 1: The energies of sequence $\mathrm{a}_0,\mathrm{a}_1$ and d_1 satisfy the following equation.

$$\|\boldsymbol{a_0}\|^2=\|\boldsymbol{a_1}\|^2+\|\boldsymbol{d_1}\|^2 \quad (8)$$

By iteration of equation (4) and (5), signal a_0 can be decomposed into a sequence $\mathrm{d}_1,\mathrm{d}_2,\cdots,\mathrm{d}_\mathrm{M},\mathrm{a}_\mathrm{M}$. The sequence contains the same information as a_0 and has the following properties.

$$\|\boldsymbol{a_0}\|^2=\|\boldsymbol{a_M}\|^2+\boldsymbol{\sum_{m=1}^{M}}\|\boldsymbol{d_m}\|^2 \quad (9)$$

The index M of a sequence $a_M=(a_{Mm})$ or $d_M=(d_{Mm})$ represents decomposition scale or decomposition level, coefficients a_{Mm} and d_{Mm} represent approximate coefficients and detail coefficients respectively, d_1 represents coefficients of the highest frequency range, d_2 represents coefficients of the next frequency range. For each measured signal the wavelet coefficients contained in $\mathrm{d}_1,\mathrm{d}_2,\cdots,\mathrm{d}_\mathrm{M}$ and a_M are arranged in a matrix as shown in Fig. (**1**), wherein, W represents the part with 0 filling.

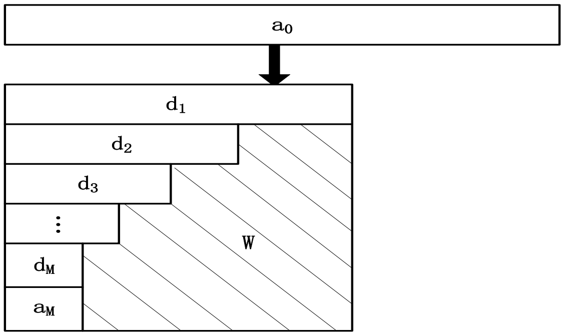

Fig. (1). Fast wavelet transform coefficients matrix of sequence a_0.

Based on above analysis, the sequence $\mathrm{d}_1,\mathrm{d}_2,\cdots,\mathrm{d}_\mathrm{M}$ and a_M can be obtained after decomposing signal a_0 with a fixed length of L by means of wavelet transform at scale M, that is, signal a_0 can be replaced by $\mathrm{d}_1,\mathrm{d}_2,\cdots,\mathrm{d}_\mathrm{M}$ and a_M, so the size of obtained wavelet coefficients matrix is $(M+1)\times(L+N-1/2)$(L is even) or $(M+1)\times(L+N/2)$(L is odd). When L is large the resulting matrix is large, and the structure of classifier becomes complex. Extracting information containing large features and merging information containing small features can reduce the dimensionality of input features, thus reducing the structure of classifier, and the classification process becomes relatively easy.

3. FEATURE EXTRACTION METHOD BASED ON WAVELET COEFFICIENTS CLUSTERING

3.1. Determination of Clustering

Fast wavelet transform is performed for K ultrasonic sampling signals. Wavelet coefficients matrix of each signal s is obtained respectively.

$$\boldsymbol{B_k=\left(b_{ij}\right)_k \qquad (k=1,2,\cdots,K)} \quad (10)$$

Wherein, $\mathrm{i}=1,2,\cdots,M+1$; $\mathrm{j}=1,2,\cdots,\frac{L+N-1}{2}$ (L is even), $\mathrm{j}=1,2,\cdots,\frac{\mathrm{L+N}}{2}$($L$ is odd), N is a singular nature number. The clustering process requires the following two results.

Theorem 2(central limit theorem) [9]: Let $\mathrm{Y_k}$ be a sequence of independent random variables, for an arbitrary $\mathrm{k}\in\mathrm{N}$ and a closed interval $[\alpha,\beta]$, $\mathrm{p}\{\mathrm{Y_k}\in[\alpha,\beta]\}=1$, and let:

$$\begin{cases}\rho_n=\sqrt{\sum\limits_{k=1}^{n}\mathrm{VY_k}}\\[2ex] \mathrm{Z_n}=\dfrac{\sum\limits_{k=1}^{n}\mathrm{Y_k}-\sum\limits_{k=1}^{n}\mathrm{EY_k}}{\rho_n}\end{cases} \quad (11)$$

Wherein, $\mathrm{n}\in\mathrm{N},\mathrm{EY_k}$ and $\mathrm{VY_k}$ respectively represent the expected value and variance of random variable Y_k , then when and only when $n\to\infty$, $\rho_n\to\infty$, Z_n obeys the standard normal distribution, that is $Z_n\overset{D}{\to}N(0,1)$.

Theorem 3 [9]: Let Y_k be an independent random variable sequence of obeying the $N(0,1)$ distributed, $\gamma\geq\mathrm{e}^2$ (e is Euler number) is a constant. Then, for any small $\varepsilon(\varepsilon>0)$, there is a natural number $N(\varepsilon)$ so that each of the random variables Z_N satisfies the following relation:

$$\boldsymbol{Z_N=\left|\left\{Y_k,Y_k\geq\sqrt{2ln\frac{N}{\gamma}},k=1,2,\cdots,N\right\}\right|} \quad (12)$$

For every natural number $\mathrm{N}\geq\mathrm{N}(\varepsilon)$, the expected value of $\mathrm{Z_N}$ satisfies the inequality:

$$\boldsymbol{\frac{\gamma}{2\sqrt{\pi lnN}}<EZ_N<(1+\varepsilon)\frac{\gamma}{2\sqrt{\pi lnN}}} \quad (13)$$

From Theorem 3 we know that when N is large enough, all independent random variables $\tilde{Z}_k(k\in N)$ [9] satisfy the following relation:

$$\boldsymbol{\tilde{Z}_k=\begin{cases}1, & Y_k\geq\sqrt{2ln\frac{N}{\gamma}}\\[2ex] 0, & Y_k<\sqrt{2ln\frac{N}{\gamma}}\end{cases}} \quad (14)$$

That is, $\tilde{\mathrm{Z}}_\mathrm{k}$ obeys binomial distribution, then $\mathrm{Z_N}$:

$$\boldsymbol{Z_N=\sum_{k=1}^{N}\tilde{Z}_k} \quad (15)$$

also obeys binomial distribution.

Using the above results, clustering can be obtained from a group of K representative signals. In different wavelet coefficients matrix $\mathrm{B_k}$, the elements in the same position can be considered as independent random variables, because they are signals sampled at different times, and the samples are independent of each other. A sequence of random variables $(b_{ij})_1, (b_{ij})_{2,}, \cdots, (b_{ij})_k$ can be formed by the elements in the same position of the wavelet coefficients matrix B_k . Thus according to the equation (11), after the random variable is transformed, a new random variable Z_{ij} can be obtained.

$$Z_{ij} = \frac{\sum_{k=1}^{K}(b_{ij})_k - \sum_{k=1}^{K} E(b_{ij})_k}{\rho_K} \tag{16}$$

Wherein, $\mathrm{E(b_{ij})_k}$ is the even value of random variable $\mathrm{(b_{ij})_k}$, and $\mathrm{Z_{ij}}$ obeys the standard normal $\mathrm{N(0,1)}$ distribution, so a new matrix can be obtained.

$$\boldsymbol{Z} = (\boldsymbol{Z_{ij}}) \tag{17}$$

The matrix is obtained by doing statistic and analyzing K wavelet coefficients matrices, therefore, it has the statistical properties of the wavelet coefficients matrix of K sample signals. Because each row of the wavelet coefficients matrix represents the size of sample signals at different scales, i.e., the magnitude of the components of different frequency regions. In other words, if the value of a row elements Z_{ij} in the matrix Z is larger, the information contained in the scale is larger, and if the value is smaller, the information contained in the scale is smaller. Thus the wavelet coefficients with larger information can be used as eigenvalues, and the wavelet coefficients with smaller information are merged, i.e., clustering. In this paper, the method of mathematical statistics was used for the clustering of wavelet coefficients. Since the size of wavelet coefficients indicates the amount of information contained, it is independent of the symbol of wavelet coefficients, therefore, wavelet coefficients are expressed by absolute values when clustering, let matrix:

$$\widetilde{\boldsymbol{B}}_{\boldsymbol{k}} = (|\boldsymbol{b_{ij}}|)_{\boldsymbol{k}} \tag{18}$$

In practice, the mathematical expectation $\mathrm{E(b_{ij})_k}$ in equation (16) can not be obtained, but average estimation can be used instead of mathematical expectation, then, the matrix in equations (16), (17) can be expressed as:

$$G = (g_{ij}) = \frac{1}{\sigma\left(R\left(\sum_{k=1}^{K}\tilde{B}_k\right)\right)}\left(\sum_{k=1}^{K}\tilde{B}_k - \sum_{k=1}^{K}\mu\left(R\left(\sum_{k=1}^{K}\tilde{B}_k\right)\right)\cdot I\right) \tag{19}$$

Wherein, R is an operator, applied to any matrix A, which represents the reduction of the dimensionality of matrix A, $\mu(A)$ and $\sigma(A)$ represent respectively the sample mean and standard variance of the elements in the matrix A, the matrix I is the same size as the matrix $\widetilde{B}_k$, but contains only element 1. According to Theorem 3, the element $\mathrm{g_{ij}}$ in G also obeys $\mathrm{N(0,1)}$ distribution, so by applying a threshold $\mathrm{T} = \sqrt{2(\mathrm{InL}/\gamma)}$ ($\gamma \geq \mathrm{e}^2$, L is the number of calculated detail coefficients) to the matrix G, the following binary matrix can be obtained.

$$\boldsymbol{G_b} = \left(\boldsymbol{\Theta}(\boldsymbol{g_{ij}} - \boldsymbol{T})\right) \tag{20}$$

For the above function $\Theta(x)$, when $x \geq 0, \Theta(x) = 1$, when $x < 0$, $\Theta(x) = 0$. Thus in the matrix $\mathrm{G_b}$, the wavelet coefficients corresponding to the element 1 are larger and contain more sample information. On the same row, the coefficients and the nearby wavelet coefficients with a value of 0 are clustered into a class. The wavelet coefficients between the two 1s are divided into two classes at the center. Clustering of wavelet coefficients in different rows, i.e., at different scales, does not overlap. Thus each clustering contains a 1, and if the entire row of the matrix $\mathrm{G_b}$ does not contain 1, the row is treated as a clustering.

3.2. Feature Extraction

The clustering process of wavelet coefficients described above shows that clustering $U_1, U_2, \cdots, U_c$ (c is the number of clustering) can be determined by a representative set of signals. In this paper, the square root of the wavelet coefficients energy is used as signal feature, thus the number of signal features is equal to the number of clusters. Thus the eigenvectors can be obtained by three steps. In the first step, fast wavelet transform is applied to sampled discrete digital signal to obtain the wavelet coefficients matrix B. In the second step, the same clustering is performed on the matrix B based on the matrix model of equation (20), and the resulting clustering is represented by $U_1, U_2, \cdots, U_c$. Row vector formed by each clustering element U_i of matrix B is represented by $r_i (i = 1,2, \cdots, c)$. In the third step, the Euclidean norm of each vector r_i is determined as a feature u_i, that is, each feature u_i is defined as the square root of energy of wavelet coefficients in the corresponding clustering U_i.

$$\boldsymbol{u_i} = \|\boldsymbol{r_i}\|_2 = \sqrt{\sum_{\boldsymbol{v}\in\boldsymbol{U_i}} \boldsymbol{V^2}} \tag{21}$$

Thus, the number c of eigenvalues of a signal s is equal to the number of clusters determined by the method in this paper. Based on above process, it can be seen that each feature $\mathrm{u_i}$ represents a set of wavelet coefficients. That is, it represents time domain information and frequency domain information of the sampled signal s, and in addition, the wavelet coefficients at different scales describe the features in a certain frequency range. According to the feature extraction process the following relation is established.

$$\|\boldsymbol{a_0}\|^2 = \sum_{\boldsymbol{i=1}}^{\boldsymbol{c}} \boldsymbol{u_i^2} \tag{22}$$

It shows that the eigenvector constructed is robust to the noise in the corresponding signal s.

4. CLASSIFICATION EXPERIMENT

Using neural network to train samples has many disadvantages such as slow training speed and easy falling into local minimum. But support vector machine (SVM) [10] can solve these problems very well. SVM has been widely

Table 1. The clustering of coefficients obtained by fast wavelet transform.

(1-1)

Scale m	Wavelet Coefficients of Signal s	Clustering Number
1	$\{d_{1(1)}, d_{1(2)}, \cdots, d_{1(512)}\}$	1
2	$\{d_{2(1)}, d_{2(21)}, \cdots, d_{2(512)}\}$	1
3	$\{d_{3(1)}, d_{3(2)}, \cdots, d_{3(13)}\}, \{d_{3(14)}, d_{3(15)}, \cdots, d_{3(512)}\}$	2
4	$\{d_{4(1)}, d_{4(2)}, \cdots, d_{4(5)}\}, \{d_{4(6)}, d_{4(7)}, \cdots, d_{4(14)}\}, \{d_{4(15)}, d_{4(16)}, d_{4(17)}\}, \{d_{4(18)}, d_{4(19)}, \cdots, d_{4(512)}\}$	4
5	$\{\mathrm{d}_{5(1)}, \mathrm{d}_{5(2)}, \mathrm{d}_{5(3)}\}, \{\mathrm{d}_{5(4)}, \mathrm{d}_{5(5)}, \cdots, \mathrm{d}_{5(8)}\}, \{d_{5(9)}, d_{5(10)}, \cdots, d_{5(512)}\}$	3
5	$\{\mathrm{a}_{5(1)}, \mathrm{a}_{5(2)}, \mathrm{a}_{5(3)}\}, \{\mathrm{a}_{5(4)}, \mathrm{a}_{5(5)}, \mathrm{a}_{5(6)}\}, \{\mathrm{a}_{5(7)}, \mathrm{a}_{5(8)}, \cdots, \mathrm{a}_{5(512)}\}$	3

used in the field of pattern recognition, and has shown excellent performance. In the absence of prior knowledge and samples, the SVM method is superior to other methods in identifying internal defects in workpieces. Next, the feature extraction method proposed in this paper was combined with SVM to classify and identify the defects existing in steel plates.

The steel plates studied in this study contain three common internal defects: delamination, porosity, and white spots [11]. The sampling frequency was 2.5MHz, 10 samples were collected from each defect, and 1024 data points near the defect waveform were intercepted. Daubechies wavelet base was adopted, the order of wavelet was 7(N=7), and the scale of wavelet transform was 5(M=5). Fast wavelet transform of each sample signal was carried out. Next, the decomposed wavelet coefficients were clustered according to the method proposed in this paper. The formula (20) is two valued, if the threshold T is too large, it contains more eigenvalues, network structure is complex, but the classification accuracy is high. Instead, if the threshold T is too small, it contains less eigenvalues, network structure is simple, but the classification accuracy is low. All things considered, T=1.5 was selected, and the clustering results are shown in Table **1**. Original signal was decomposed into 5 layers by discrete wavelet transform. Thus according to the clustering method in 3.1, matrix A was clustered into 14 classes $\mathrm{U}_1, \mathrm{U}_2, \cdots, \mathrm{U}_{14}$. In Table **1**, the signal sequence corresponding to scale 4 contains four clusters, and the signal sequence corresponding to scale 3 contains two clusters, indicating that the signal sequence corresponding to scale 4 carries more useful information than the signal sequence corresponding to scale 3. In this experiment, there were 3 * 10 = 30 learning samples, so there were 30 sets of clustered wavelet coefficients, and the dimension of the feature vector of each sample was 14.

Ten test samples were obtained from each kind of defect, so thirty test samples were collected. Using SVM as classifier and radial basis function as kernel function of SVM classifier, one-to-one classification method was used, and the classification accuracy was as high as 96%. The wavelet coefficients of the fourth scale were used as the extracted features, and the same recognizer was used. The accurate rate was as high as 90%. The wavelet coefficients of the fifth scale were used as the extracted features, and the same recognizer was used. The accurate rate was as high as 83%. The proposed feature extraction method can retain all the wavelet decomposition coefficients and does not lose any information. In addition, the clustering of wavelet coefficients can reduce the dimensionality of data. So it can simplify the recognition process of recognizer. Therefore, the proposed feature extraction method can improve the accuracy of defect recognition, and it is effective and reliable for the defect recognition of ultrasonic nondestructive testing.

CONCLUSION

In this paper, the identification of defects in steel plate was taken as an example. Firstly, the fast wavelet transform of sample signals was used to obtain the wavelet coefficients matrix. Secondly, the wavelet coefficients containing more sample information in the matrix were extracted by the clustering method of probability and statistics, and the wavelet coefficients containing less sample information were merged. And finally, the wavelet coefficients energy value of each cluster was calculated and used as the input feature vector of the SVM classifier. Each feature represents a set of wavelet coefficients, which can preserve the information of particular problem of measurement signal, and greatly reduce the dimension of input vector. Proposed feature extraction method makes the pattern recognition process easier and improves the classification accuracy.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was financially supported by project of the science and technology development fund of Tianjin City (20110831).

REFERENCES

[1] Xia Jizhen. Ultrasonic Nondestructive Testing Technology, CHN: Guangdong, pp. 235-242, 2009.

[2] C.K. Chui. An Introduction to Wavelets, USA: Boston, pp.197-212, 1992.

[3] H. Hatanaka, Y. Kawano, N. Ido, M. Hato, and M. Tagami. "Ultrasonic testing with advanced signal processing for concrete structures", NDT&E, vol. 20, pp. 115-124, Feb. 2005.

[4] R. Polikar, L. Udpa, S. Udpa, and T. Taylor, "Frequency invariant classification of ultrasonic weld inspection signals", *IEEE Trans. On Ultra. Ferro. and Freq. Control*, vol. 45, pp. 614-625, Mar. 1998.

[5] Tang Baoping, and Qin Shuren, Tan Shanwen. "The Signals Analysis System Based on Fast Wavelet Transform", *Journal of Chongqing University (Natural Science Edition)*. vol.24, pp. 5-8, Jan. 2001.

[6] Li Guozheng, Wang Meng, and Zeng Huajun. "An introduction to support vector and other kernel-based learning methods", *CHN: Beijing*, pp. 202-210, 2004.

[7] Zhang Xuegong. "Statistical learning theory and support vector machines", *Acta Automatica Sinica*. vol.26, pp. 32-41, Jan. 2000.

[8] Daubechies I. "The wavelets transform time-frequency localization and signal analysis", *IEEE Trans on Information Theory*, vol. 36, pp. 961-1005, May. 1990.

[9] He Shengwu. Probability Theory and Mathematical Statistics. *CHI: Beijing*, pp. 143-150, 1992.

[10] Li Guozheng, Wang Meng, Zeng Huajun. An introduction to support vector machine (SVM), *CHN: Beijing*, pp. 210-218, 2004.

[11] Cui Hong. "Ultrasonic Inspection for Steel Plate and the Defects Discrimination", *Liugang technology*, vol. 2, pp. 45-48, Feb. 2008.

Received: May 26, 2015 Revised: July 14, 2015 Accepted: August 10, 2015