An Optimized VTCR Feature Dimensionality Reduction Algorithm Based on Information Entropy

Shaohao Mou, Weikuan Jia*, Yuyu Tian, Yuanjie Zheng, Yanna Zhao

***Abstract*—As the basic research applied in pattern recognition, machine leaning, data mining and other fields, the main purpose of feature extraction is to achieve low loss of data dimensionality reduction. Among all the dimensionality reduction algorithm, the classical statistical theory is the most widely used, the feature variance total contribution ratio (VTCR) is mostly used to measure the effect of evaluation criteria for feature extraction. Traditional VTCR only focuses on the nature of the samples' correlation matrix eigenvalue but not the information measurement, resulting in large loss of information for feature extraction. Shannon information entropy is introduced into feature extraction algorithm, the generalized class probability and the class information function are defined, the contributive ratio for VTCR is improved. Finally, the dimensions of feature extraction are determined by calculating the accumulate information ratio (AIR), which could achieve good evaluation in respect of information theory. By combining the new methods with principal component analysis (PCA) and factor analysis (FA) respectively, an optimized VTCR feature dimensionality reduction algorithm based on information entropy is established; the number of feature dimensions extracted is calculated by AIR. By the experiment, the results show that, the low-dimensional data has more interpretability, and the new algorithm has higher compression ratio.**

***Index Terms*—Feature Extraction; Variance Total Contribution Ratio; Shannon Entropy; Accumulate Information Ratio**

Manuscript received October 13, 2018. This work is supported by Focus on Research and Development Plan in Shandong Province (2019GNC106115); China Postdoctoral Science Foundation (No.: 2018M630797); project of Shandong Province Higher Educational Science and Technology Program (J18KA308); National Nature Science Foundation of China (No.: 31571571, 61572300); Taishan Scholar Program of Shandong Province of China (No.: TSHW201502038).

S.H. Mou is with School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China

W. K. Jia is with School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (Corresponding author, Tel: +86-531-86190755; Email: jwk_1982@163.com)

Y. Y. Tian is with School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China

Y. N. Zhao is with Key Lab of Intelligent Computing & Information Security in Universities of Shandong, Shandong Normal University, Jinan 250358, China

Y. J. Zheng is with Institute of Life Sciences, Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology & Key Lab of Intelligent Information Processing, Shandong Normal University, Jinan 250014, China

I. Introduction

With the advent of information technology, data sets update and grow faster, and number of data dimensions become higher and unstructured. Although data obtained through these information systems contains a wealth of information, much available information also increase the difficulty of effective data analysis. For instance, data redundancy results in overwhelmed useful information, the correlation between the features of the data causes repeated information, and so on. So how to make use of these huge volumes of data, analyze, extract useful information and exclude the influence of unrelated or repeated factors, to save the storage space and computing time. Feature extraction theory is that extracting effective and reasonable simple data from the huge volumes data while keeping the data information complete, which is to reduce the feature dimensions without affecting the problem solving as much as possible in order to satisfy the need of storage, computation and recognition [1-4].

Classical statistical analysis methods are commonly feature extraction algorithms, which are usually judged by the value of the variance total contribution ratio (VTCR) [5]. In many application fields, there are many mature methods such as principal component analysis (PCA) [6], independent component analysis (ICA) [7], factor analysis (FA) [8], correlation analysis [9], cluster analysis [10], linear discriminant analysis (LDA) [11], etc. these methods have achieved good results in dealing with linear problems. However, nonlinear problems are ubiquitous, and how to improve these classical algorithms to deal with nonlinear problems, it is the current research hotspot. The most studied Kernel-based improvements, such as K-PCA [12, 13], K-ICA [14], two-dimensional principal component analysis [15], and so on, they have obtained good effect. For these algorithms, VTCR is calculated based on the correlation matrix eigenvalues of data sample to measure the quality of feature extraction and to determine the number of extracted features. Although this method is simple to operate, but the eigenvalues of relation matrix hardly covers information measure, in other words, VTCR can't really evaluate the efficacy of dimension reduction algorithm from the angle of information.

in the information theory, information entropy (Shannon entropy) is used to measure the amount of information, it establishing the scientific foundation for the modern

information theory. The information entropy is proposed by the American institute of electrical engineers Shannon CE in 1948 [16], entropy is considered as a measure of "uncertainty" of a random incident or the amount of information, as a variable the uncertainly larger, the entropy larger, the greater the amount of information contained. The information entropy theory is used for feature extraction, it has got many achievements, some algorithms are directly be used for feature extraction [17], some are integrated with other algorithms [18].

In this study, PCA and FA are selected as typical example of VTCR. PCA is a method to represent the original multiple variables with several comprehensive factors and make the comprehensive factors reflect as much information of the original variables as possible, the comprehensive factors do not relate to each other so as to reduce dimensions. FA is one kind of analysis method to change many variables into the several integrated variables, it concentrates the information of the system's numerous original indexes and saves to the factors, it can also adjust the amount of the information by controlling the number of the factors, according to the precision that the actual problems need. FA can be seen as a further promotion of the PCA, to a certain extent, the data after dimension reduction by FA include more original information than by PCA.

Aiming at VTCR limitations, in this paper, some concepts are defined, such as generalized class probability (GCP), class information function (CIF), accumulate information ratio (AIR) based on the eigenvalue of sample correlation matrix combining with the Shannon entropy theory. The feature dimension is extracted by calculating AIR and measure the quality of feature extraction. Further, an optimized VTCR feature dimensionality reduction algorithm based on information entropy is established, which calculated AIR as the original data contained by the extracted feature, to determine the principal components or factor numbers, in order to achieve dimension reduction. New algorithm evaluates from the angle of information theory has more original information, by experiment, the results show that, the low-dimensional data has more interpretability, and the new algorithm has higher compression ratio.

II. Information Entropy and Variance Total Contribution Ratio

A. Shannon Information Entropy

The amount of information is the core of information theory, and it is the basic starting point of measure information, the obtained information is regarded as indicator to eliminate uncertainty. The amount of information describes the size of the eliminated uncertainties, and the probability distribution describes the size of uncertainty of random events.

Suppose discrete random variable X , there are *n* possible values for $a_1, a_2, \cdots, a_n$, the probability of each result is respectively $p_1, p_2, \cdots, p_n$, the probability space X can be expressed as

$$[X \bullet P]: \begin{cases} X: a_1, a_2, \cdots, a_n, \\ P: p_1, p_2, \cdots, p_n, \end{cases} \quad \sum_{i=1}^{n} p_i = 1 \qquad (1)$$

Where, $P(a_i) = p_i$ is the probability of event $\{X = a_i\}$ occurred, and $0 \le p_i \le 1$. Since $[X \bullet P]$ completely describes the characteristic of the discrete information source represented by X , it is called the information source space for the information source *X*.

Definition:

Information function

$$I(a_i) = -\log p_i, i = 1,2,\cdots,n \qquad (2)$$

It represents the uncertainty of the event $\{X = a_i\}$ before the event $\{X = a_i\}$ occurs and the contained information of the event $\{X = a_i\}$ after the event $\{X = a_i\}$ occurs, which is also called self-information for a_i. Shannon makes the statistical average of information source space $[X \bullet P]$ in the information function

$$H(X) = H(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_i \log p_i \qquad (3)$$

as the measurement of the uncertainty for information source X , what we call the Information Entropy or Shannon Entropy. As $H(X)$ is larger, the uncertainty of information source X larger, and the obtained amount of information more.

B. Variance Total Contribution Ratio

First, the original data is normalized so as to eliminate the differences of index distribution, which can not only avoid repetition of the information but also overcome subjective factor in weight determination, ensuring the utilization quality from data headstream.

Suppose there are *n* variables in the original sample, denoted by $X = x_1, x_2, \cdots, x_n$, through the orthogonal transformation, integrated into n comprehensive variables, namely:

$$\begin{cases} y_1 = c_{11}x_1 + c_{12}x_2 + \cdots + c_{1p}x_n \\ y_2 = c_{21}x_1 + c_{22}x_2 + \cdots + c_{2p}x_n \\ \cdots \\ y_n = c_{n1}x_1 + c_{n2}x_2 + \cdots + c_{nn}x_n \end{cases} \qquad (4)$$

and they meet the following equation:

$$c_{k1}^2 + c_{k2}^2 + \cdots + c_{kn}^2 = 1 \;\; k = 1,2,\cdots,n \qquad (5)$$

In which y_i and y_j ($i \ne j$, $i, j = 1,2,\cdots,p$) are independent, thus *X* variance transferred to the comprehensive variables $y_1, y_2, \cdots, y_n$.

By the correlation coefficient matrix R of sample *X*, Jacobi method is used, solution of the roots of characteristic equation

$$|\lambda I - R| = 0 \qquad (6)$$

n non-negative eigenvalues $\lambda_i (i = 1,2,\cdots,n)$ of the correlation coefficient matrix of the sample has been got, and then sorting, there are $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge 0$. From *n* comprehensive variables extract front m features, the proportion of the variance of the former *m* principal

components taking up all of the variances can be defined as VTCR, denote by α:

$$\alpha = \sum_{i=1}^{m} \lambda_i \Big/ \sum_{1}^{n} \lambda_i \tag{7}$$

In practice applications, *m* is determined by the value of α, which plays the role of dimension reduction, and α represents the original data information included in feature extraction.

C. Basic Principle of PCA and FA

PCA algorithm uses the VTCR to determine the number of extracted principal components, i.e. *m* principal components. When the value of α approximates 1, therefore, the principal components of $y_1, y_2, \cdots, y_m$ basically embrace information of the original variables $x_1, x_2, \cdots, x_n$, and the number of variables has reduced from *n* to *m*, while these *m* variables contain the major original information and can reduce dimensions. Previously described VTCR is the PCA algorithm theory.

The basic philosophy of FA is to divide the observation variables into several classes, make the ones which are related close in the same class, the relativity between the variables of different classes is lower, then each class of variables represents a basic structure in fact, that is the public factor. Then we can discover each variable's best subset from numerous factors, describe the multivariable systems results and the influence on the system of the various factor from the information included in the subsets.

FA algorithm and the PCA algorithm are slightly different, supposes the observable random vector $X_i = x_1, x_2, \cdots, x_n$ but the unobservable vector $F_j = F_1, F_2, \cdots, F_m$

$$X_i = \sum_{j=1}^{m} a_{ij} F_j + c_i \varepsilon_i \quad i = 1,2,\cdots,n;\ j = 1,2,\cdots m \tag{8}$$

In this formula, $n > m$, a_{ij} is the factor loading, represents the correlation coefficient of the *i-th* variable and the *j-th* factor, reflects the importance of the *i-th* variable to the *j-th* factor, F is called public factor, they are the factors which appear in the expression of each original observation variable, and are mutually independent unobservable theoretical variable. c_j represents the load of the unique factor, ε_i affects the unique factor of X_i. The basic question of FA is to decide the factor loading by the correlation coefficient between variables. Supposes ***A*** is the factor loading matrix, namely,

$$A = (a_{ij})_{n \times m} \tag{9}$$

Determine the number of factors extracted according to the value of α, further get $\boldsymbol{A}$, calculate the synthesis score of the factors so as to achieve dimensionality reduction.

III. Accumulate Information Ratio

A. Generalized Class Probability

Definition 1

Suppose the sample has *n* characteristics, can be calculated for each feature corresponding eigenvalue and eigenvector. For eigenvalue λ_i, we make the following transformation, define generalized class probability ρ_i, using the form of probability to describe the eigenvalue of the sample

$$\rho_i = 1 - \lambda_i \Big/ \sum_{i=1}^{n} \lambda_i \tag{10}$$

so $0 \le \rho_i \le 1, \sum_{i=1}^{n} \rho_i = 1$, ρ_i represents the eigenvalue λ_i of feature is larger, the smaller probability of this feature is weed out.

B. Class Information Function

Definition 2

Combining the definition of information function of the information theory, by the generalized class probability, we further define Class Information Function $I(\lambda_i)$, also known as the Class amount of self-information

$$I(\lambda_i) = -\log \rho_i \tag{11}$$

$I(\lambda_i)$ represents the amount of information for λ_i, as λ_i is larger, ρ_i smaller, and the $I(\lambda_i)$ larger, λ_i has more loading information, so this according with the inherent characteristics of the information function.

Loading information is sorted through the class information function.

$$I(\lambda_1) \ge I(\lambda_2) \ge \cdots \ge I(\lambda_n) \ge 0 \tag{12}$$

C. Class Information Ratio

The class information ratio is used to represent the amount of information carried by λ_i

$$IR(\lambda_i) = I(\lambda_i) \Big/ \sum_{i=1}^{n} I(\lambda_i) \tag{13}$$

$IR(\lambda_i)$ represents the contribution rate of information for λ_i, and sorted by the contribution rate of information

$$IR(\lambda_1) \ge IR(\lambda_2) \ge \cdots \ge IR(\lambda_n) \ge 0 \tag{14}$$

and satisfy

$$IR(\lambda_1) + IR(\lambda_2) + \cdots + IR(\lambda_n) = 1 \tag{15}$$

D. Accumulated Information Ratio (AIR)

Definition 3

Here, the accumulated information ratio *AIR* is defined.

$$AIR = \sum_{i=1}^{m} I(\lambda_i) \Big/ \sum_{i=1}^{n} I(\lambda_i) \tag{16}$$

In practice, *m* is valued by *AIR*, while accumulated information represents the original amount of information of feature extraction data. *AIR* is obtained through information entropy, so it takes information measurement into account and evaluates feature extraction effect in respect of information theory.

E. Compression Ratio

In order to better describe the performance of data feature extract algorithm, and to show more intuitive the ability of dimension reduction algorithm, the concept of compression ratio η is introduced.

Compression ratio is defined as

$$\eta = \frac{n-m}{n} \times 100\% \tag{17}$$

Namely, the compression ratio η represents the degree of compression data dimension.

IV. Optimizing Variance Total Contribution Ratio Algorithm Based on Information Entropy

According to the aforementioned theory, PCA algorithm is considered in information function to establish PCA algorithm based on information entropy. The selection mechanism of the principal component is changed by defined accumulated information ratio, calculate the loading matrix, and further propose the PCA feature extraction algorithm based on the information entropy. Suppose the original data X is $s \times n$ matrix, containing s samples with n dimensional feature.

The basic steps of the new algorithm are as follows:

Step 1 Normalize the original data, the transformed matrix is still denoted by X;

Step 2 Find the X correlation coefficient matrix R, and find the correlation orthogonal matrix U by R;

Step 3 In Jacobi method, find the eigenvalue λ_i and eigenvector u_i of the matrix $\boldsymbol{R}$ through the characteristic equation $|\lambda I - R| = 0$;

Step 4 Calculate generalized class probability ρ_i, class information function $I(\lambda_i)$, accumulated information ratio *AIR*;

Step 5 Determine m extracted principal components by the value of *AIR*;

Step 6 Decompose the eigenvector u_i by extracted number of principal components.

Step 7 Normalize the decomposed eigenvector, write out the principal component loading matrix $(a_{ij})_{s \times m}$;

Step 8 Writing the principal component loading matrix according to the practical problem, calculate the principal component score, the algorithm terminates.

Similarly, we combining information function with FA algorithm, first established FA algorithm based on information entropy. Also, according to the AIR to determine the number of factors is extracted. The two algorithms have slightly different at step 8, the principal component loading matrix not need to rotate, if rotate, the results are not principal component. But the factor analyses, according to the practical problems, to determine the loading matrix whether need to rotate, get the factor with stronger interpretability and more meaningful.

By the above algorithm, the dimension of the original data is compressed from *n* into *m*. With the concept of accumulated information ratio defined by information function, this algorithm covers not only the characteristics of eigenvalues but also the information measurement from *m* selection mechanism. Finally, we got the low dimensional data containing most of the original data information to achieve the goal of feature dimension reduction.

V. Experiment

In order to test the reliability of the new algorithm, two groups of experiment are arranged. The experimental data sets are from actual production data and UCI machine leaning standard data sets. The actual production data, wheat blossom midge in Guanzhong area is selected [19]; another is the waveform generator set, which from UCI machine learning standard data sets [20].

A. Test 1

In actual production, we have known data, using the weather factors to forecast the occurrence degree of the wheat blossom midge in Guanzhong area. Here we choose the data of 60 samples from 1941 to 2000 as the study object, which with 14 feature variables and 1 dependent variables.

According to the section 3 established optimization algorithm steps, we calculated the X correlation coefficient matrix $\boldsymbol{R}$ and found the eigenvalue λ_i and eigenvector u_i in Jacobi method. Variance total contribution ratio (VTCR) is measured to further calculate the generalized class probability ρ_i, class information function $I(\lambda_i)$, accumulated information ratio *AIR*. The results are listed in Table 1.

TABLE I
THE RESULTS OF λ_i, *TCR*, ρ_i, $I(\lambda_i)$, *AIR* FOR TEST 1

NO.	λ_i	*VTCR* (%)	ρ_i	$I(\lambda_i)$	*AIR*(%)
1	2.9724	23.37	0.7663	0.2661	24.82
2	1.8264	37.72	0.8564	0.1550	39.28
3	1.7482	51.47	0.8626	0.1478	53.06
4	1.4665	62.99	0.8847	0.1225	64.49
5	1.1636	72.14	0.9085	0.0959	73.44
6	0.8639	78.93	0.9321	0.0705	80.01
7	0.6957	84.40	0.9453	0.0560	85.23
8	0.5275	88.55	0.9585	0.0424	89.19
9	0.4362	91.98	0.9657	0.0349	92.44
10	0.3223	94.51	0.9747	0.0257	95.01
11	0.2643	96.59	0.9792	0.0210	96.80
12	0.1950	98.12	0.9847	0.0154	98.24
13	0.1405	99.23	0.9890	0.0111	99.27
14	0.0985	100.00	0.9923	0.0078	100.00

If the amount of information can reach 85%, the solution will not be influenced. Table 1 show that, evaluation from the aspect of information theory, by *AIR* standard, we extract 7 principal components (factors), which can be amounted to 85.23% of the original information. However, by *VTCR* standard, we need extract 8 principal components (factors), to make the amount of information more than 85%.

From compression ratio, *VTCR* standard compression ratio is 42.86%, the compression ratio of new algorithm is 50%, more conducive to the further processing of data.

B. Test 2

As data is known, the waveform generator set of the UCI standard data set, we select 5000 samples from data as research objects and use $x_1 \sim x_{40}$ to represent 40 characteristic variables of the raw data. The simulation results of ionosphere radar data set list in table 2. In order to save paper space, here only lists the key data about determine to extract feature, the remaining data is not listed in the article.

Table 2 show that, according to *VTCR* standard, we need to extract 26 principal components (factors); According to AIR standard, we only need to extract 24 principal components (factors), two less than that by *VTCR* standard but acquire more precise results from amount of information level.

TABLE II
THE RESULTS OF λ_i, *TCR*, ρ_i, $I(\lambda_i)$, *AIR* FOR TEST 2

NO.	λ_i	*VTCR* (%)	ρ_i	$I(\lambda_i)$	*AIR* (%)
21	0.9327	76.1316	0.9767	0.0340	78.6570
22	0.9160	78.4216	0.9771	0.0334	80.7488
23	0.9068	80.6886	0.9773	0.0331	82.8194
24	0.8770	82.8810	0.9781	0.0320	85.1212
25	0.8534	84.7146	0.9787	0.0291	86.9434
26	0.6828	86.7216	0.9829	0.0248	88.1981
27	0.6655	88.3853	0.9834	0.0232	89.6505
28	0.4825	89.5916	0.9879	0.0175	90.7463
29	0.4818	90.7960	0.9880	0.0165	91.7780
30	0.4334	91.8795	0.9892	0.0157	92.7617

C. Results

Form table 1 and 2, if the same numbers of components (factors) are extracted, from the perspective of information content, the results of *AIR* standards more precise than the standard *VTCR*. If under the same content of information, the components (factors) are extracted by *AIR*, its numbers are fewer than extracted by *VTCR*. By *AIR* method, the low-dimensional data has more interpretability, and the new algorithm has higher compression ratio.

By the two groups of experimental results show that the new evaluation criterion can extract the fewer features, and the low-dimensional data has a more appropriate explanation. The dimension of two data sets is 14d, 40d, the dimension reduction effect of new algorithm has been reflected. If in the practical application, thousands and thousands of dimensional data, even higher dimensions, the effect will be more obvious. As a new feature extraction and evaluation mechanism, the new method is effective and feasible.

VI. CONCLUSION

In this paper, the proposed algorithm theory is variance total contribution ratio on the basis of optimized information entropy, which is called accumulated information ratio. In order to verify the reliability of the new feature extraction mechanism, the PCA feature extraction and the FA feature extraction algorithm based on information entropy are established by PCA or FA algorithm in combination, making the new theory practice in the PCA and FA. By the experiment's analysis, the results show that, the consideration of new selection mechanism is more comprehensively. When it is used to describe the amount of information contained, it shows the superiority of the new mechanism.

The new proposed new selection mechanism for feature extraction algorithm that changed the way to determine the number of extracted features simply relied on variance total contribution ratio. New theory fully takes into account the nature of the sample eigenvalues and the measurement of information and proves an excellent choice and evaluation mechanisms. The accumulated information ratio as evaluation criterion can better depict the compressed degree of information, which is worthy of further promotion.

REFERENCES

[1] S. F. Ding, H. Zhu, W. K. Jia, & C. Y. Su, "A survey on feature extraction for pattern recognition," *Artif. Intell. Rev.*, vol.17, no.3, pp. 169-180, 2012.

[2] Y.Y. Ma, & L.P. Zhu, "A review on dimension reduction," *Int. Stat. Rev*., vol. 81, no. 1, pp.134-150, 2013.

[3] K. Aihara, K. Abe, & E. Ishiwata. "Preconditioned IDRStab algorithms for solving nonsymmetric linear systems, " *International Journal of Applied Mathematics*, vol. 45, no. 3, pp. 164-174, 2015.

[4] L. L. Zhu, Y. Pan, M. K. Jamil, & W. Gao, "Boosting based ontology sparse vector computation approach" *Engineering Letters*, v 25, n 4, p 406-415, 2017.

[5] W. K. Jia, S. F. Ding, X. Z. Xu, C. Y. Su, & Z. Z. Shi. "Factor analysis feature extraction algorithm based on Shannon entropy. *Pattern Recognition and Artificial Intelligence*, vol. 24, no. 3, pp. 327-331, 2011. (in Chinese)

[6] L. I. Kuncheva, & W. J. Faithfull. "PCA feature extraction for change detection in multidimensional unlabeled data." *IEEE T. Neur. Net. Lear*, vol.25, no.1, pp.69-80, 2014.

[7] O. W. Kwon, T. W. Lee. "Phoneme recognition using ICA-based feature extraction and transformation," *Signal Processing*, vol. 84, no. 6, pp. 1005-1019, 2014.

[8] W. Husin, W. Zakiyatussariroh, M. S. Zainol, & N. M. Ramli, "Common factor model with multiple trends for forecasting short term mortality," *Engineering Letters*, vol. 24, no. 1, 98-105, 2016.

[9] Z. Zhang, M. B. Zhao, & T. W. S. Chow. "Binary and multi-class group sparse canonical correlation analysis for feature extraction and classification." *IEEE T Knowl. Data En.*, vol. 25, no. 10, pp. 2192-2205, 2013.

[10] K. S. Lin, & C. F. Chen, "Cluster analysis of genome-wide expression data for feature extraction." *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3327-3335, 2009.

[11] S. S. Arifin, and M. H. Muljono, "A Model of indonesian dynamic visemes from facial motion capture database using a clustering-based approach." *IAENG International Journal of Computer Science*, vol. 44, no.1, pp. 41-51, 2017.

[12] J. Li, X. L. Li, & D. C. Tao, "KPCA for semantic object extraction in images." *Pattern Recogn.*, vol. 41, no. 10, pp. 3244-3250, 2008.

[13] K. Chougdali, Z. Elkhadir, & M. Benattou, "Intrusion detection system using PCA and kernel PCA methods," *IAENG International Journal of Computer Science*, vol. 43, no.1, pp. 72-79, 2016.

[14] A. B. Musa. "A comparison of ℓ1-regularizion, PCA, KPCA and ICA for dimensionality reduction in logistic regression." *Int. J. Mach Learn Cyb.*, vol. 5, no. 6, pp. 1-13, 2013.

[15] Y. Choi, S. Ozawa, & M. Lee, "Incremental two-dimensional kernel principal component analysis." *Neurocomputing*, vol. 134, pp. 280–288, 2014.

[16] C. E. Shannon. "A mathematical theory of communication." *Bell Syst. Tech. J.*, vol. 27, no. 3, pp.379-423, 1948.

[17] H. Q. Wang, & P. Chen, "A feature extraction method based on information theory for fault diagnosis of reciprocating machinery." *Sensors*, vol. 9, no. 4, pp. 2415-2436, 2009.
[18] S. F. Ding, & Z. Z. Shi. "Studies on incidence pattern recognition based on information entropy." *J. Inf. Sci.*, vol. 31, no. 6, pp. 497-501, 2005.
[19] Y. M. Zhang. "The application of artificial neural network in the forecasting of wheat midge." Master thesis, Northwest A&F University, 2003. (in Chinese)
[20] DatabaseGenerator (Version2) (2016, October 10) http://www.ics.uci.edu/~mlearn/databases/Waveform