ICIC Express Letters
Volume **10**, Number **5**, May **2016**

 ISSN 1881-803X
pp. **1219–1224**

COMBINATION OF INTERACTIVE FEATURE EXTRACTION USING QUADRATIC MAP AND FEATURE SELECTION USING GENETIC ALGORITHM

Jing Li[1,2], Cunfang Zheng[2] and Wenxue Hong[2]

[1]School of Science
[2]School of Electrical Engineering
Yanshan University
No. 438, Hebei Avenue, Haigang District, Qinhuangdao 066004, P. R. China
01016888@sina.com

Received October 2015; accepted January 2016

Abstract. *The graphical representation of multi-dimensional data is a simplest method of the high dimensional data visualization. The interactive barycentre graphical feature is proposed based on the radar plot of multi-dimensional data. The radar plot is involved with the feature order, which leads that interactive feature is involved with the feature order. The quadratic map is used to obtain interactive features of the all feature orders. The discriminating feature selection method is proposed based on the improved genetic algorithm (GA). For some UCI data set such as wine, breast cancer and diabetes, the obtained best classification error of discriminating interactive feature of radar plot is 0.0%, 1.6% and 20.7%, which is very promising compared to the previously reported classification methods, and is superior to that of traditional feature extraction method.*
Keywords: Data visualization, Graphical representation, Feature extraction, Quadratic map, Feature selection, Genetic algorithm

1. **Introduction.** Feature extraction and classification are the key when extracting meaningful information from the complicated sensor data. An ideal feature extraction method should produce a good representation of data [1], which makes subsequent classification easier. A general idea is to extract feature which has the most discriminant ability, such as fisher linear discriminant analysis (FLDA), and generalized discriminant analysis (GDA) [2].

Here the problem of feature extraction from the perspective of data visualization is reviewed. The graphical representation of multi-dimensional data is a simplest method of the high dimensional data visualization. It contains parallel coordinate, radar plot, scatter plots, Chernoff face, constellation graph and so on [3]. A radar plot graphical representation based on high dimensional data samples realizes the visualization of high-dimensional data, and at the same time it gives visualization information, which makes it easier for human to classify and understand the data [4].

Duin and his partners studied the representation method based on dissimilarity [5]. A paper published in *Nature* pointed out [6]: "It analyses selected parts of its visual environment according to a small number of pattern parameters such as size, color or contour orientation, and stores particular parameter values".

So we proposed interactive barycentre graphical features extraction method based on graphical representation of multivariate data and biology theory, which has the ability of visualization and classifying. Now the high-dimensional data first will be represented as radar plot. Then the interactive graphical features are extracted from radar plot. Finally the features are put into the classifier. Due to the difference in data sorting, the features of radar plot are affected by the shape of radar plot, which is affected by the data feature order. Based on the quadratic map [7], all the graphic features of all radar plot of all the

data feature order are calculated. So we proposed interactive features based on quadratic map.

Due to the high dimension of interactive features, the feature selection methods should be used to select the discriminating interactive feature for classification. The genetic algorithm, particle swarm optimization or differential evolution algorithm are usual feature selection methods [8]. The interactive features subset are selected based on genetic algorithm. The classifier uses the traditional classifier.

So the combination of interactive feature extraction using quadratic map and feature selection using genetic algorithm is proposed for classifying. The experimental results of some UCI data sets such as wine, breast cancer and diabetes confirmed our thoughts.

2. Interactive Feature Extractions Using Quadratic Map.

2.1. Data visualization and interactive feature extractions.

The radar plot is one of the most commonly used in graphical representation of multivariate data method. Radar plot is simple and subjective. Suppose that the data are normalized to the uniform distribution ranging from 0 to 1. A closed irregular polygon with multiple triangles is formed in a 2-dimensional plane of radar plot. Each triangle is composed of adjacent variable point and original point. Each triangle of radar plot has a barycentre. The barycentre feature G is defined as the distance from the origin point to barycentre point of each triangle of radar plot. The sample with d-dimensional feature produces the d-dimensional barycentre features. The formula is as follows:

$$G_i = f(r_i, r_{i+1}) = \frac{\sqrt{r_i^2 + r_{i+1}^2 + 2r_i r_{i+1}\cos\omega_i}}{3}, \quad \omega_i = 2\pi/d, \quad i = 1, 2, \cdots, d \tag{1}$$

So Formula (1) meets the commutative law $f(r_i, r_{i+1}) = f(r_{i+1}, r_i)$. Specially, when $i = d$, $i + 1 = d + 1$ is the $d + 1$ dimension. Because the radar plot is variable along the circumference of a circle, we think the $d + 1$ dimension is 1-dimension. Obviously, the feature order will influence the shape of radar plot, which affects barycentre features and the classification performance eventually. So we fused interactive barycentre features of radar plot under the all feature order.

2.2. Quadratic mapping.

Quadratic mapping refers to a d dimension data $x = \{x_1, \ldots, x_i, x_{i+1}, \ldots, x_d\}$ is mapped to a new data with $d\,(d+3)/2$ dimension spaces, as shown in Formula (2).

$$Y = \begin{bmatrix} x_1 & x_2 & \cdots & x_d \\ x_1x_1 & x_1x_2 & \cdots & x_1x_d \\ x_2x_1 & x_2x_2 & \cdots & x_2x_d \\ \vdots & \vdots & \cdots & \vdots \\ x_dx_1 & x_dx_2 & \cdots & x_dx_d \end{bmatrix} \tag{2}$$

Obviously $x_ix_j = x_jx_i$, so Formula (2) the lower triangular matrix is ignored. If we replace x_ix_j with the general expression of a two variable function $f(x_i, x_j)$, then we are more likely to set up the contact with Formula (1). Interactive feature $f(r_i, r_{i+1})$ can be as a special kind of implementation of $f(x_i, x_j)$. So the interactive barycentre features of radar plot under all the possible features order are the elements of Y. Then the problem of feature order affecting interactive barycentre feature mentioned in the former section can be solved. It is transformed into a feature selection problem. First a set of d dimension data according to the interactive feature extraction Formula (1) by quadratic mapping ascends to a $d(d + 3)/2$ dimension data, and then the d dimension interactive barycentre features are chosen from $d(d+3)/2$ dimension of high dimensional space, and last we hope the d dimension interactive barycentre features with classification identification ability.

3. **Feature Selection Using Genetic Algorithm.** Genetic algorithm (GA) is an adaptive optimization search algorithm [9,10], which directly simulates Darwin's theory of natural selection and genetics in the biological system. Based on Darwin's principle of survival of the fittest, after a series of iterative genetic operation, the GA optimization solutions are obtained. The chromosome with highest fitness of the final population is corresponding to the acceptable solution of the problem.

3.1. **Chromosome design.** When applying GA to feature selection, the key question is how to code the solution as chromosome. We chose an efficient encoding way: integer encoding. The coding scheme requires that each gene in each chromosome is different from others, and each gene must be a positive integer less than or equal to the maximum number of dimensions $d(d+3)/2$, and the gene length equal to the selected dimension number, such as d, and the genes order is arbitrary. The initial population is composed randomly by multiple chromosomes with randomly selected features, and each chromosome is a kind of feature selection.

3.2. **Fitness function.** Fitness function indicates the quality of the chromosome. The commonly used fitness function includes distance measurement, information entropy, correlation measurement, consistency and classifier accuracy, etc. Here fitness function is designed as the classification accuracy of interactive barycentre features selected by each chromosome. The classification performance is the 10-fold cross validation (10 CV) accuracy of the classifier. The classifier includes the linear discriminate analysis (LDA) classifier, nearest neighbor classifier and K neighbor classifier (KNN), and support vector machine (SVM). The higher the fitness function values, the better the interactive barycentre features chosen by chromosomes.

3.3. **Genetic manipulation and repair operator.** Fitness function evaluates the quality of chromosome. The standard genetic algorithm operations, such as selection, crossover and mutation are applied to the population. With the roulette wheel selection operation, chromosomes are selected in a higher probability into mating pool when the fitness function of chromosomes is higher. General crossover operations with single point or multipoint crossover exchange the genes on the two chromosomes, and mutation operation changes a gene and certain genes on chromosome. However, for feature selection problem, the standard genetic algorithm operation may produce infeasible solutions, which could not meet the requirements of the coding scheme of feature selection. We proposed the post-processing method suitable for our problem, called repair operator, which turns all infeasible solutions generated by standard genetic method into feasible ones and estimates the value of fitness function of each legitimate chromosome after repairmen. The proposed repair operator includes four steps: first step is the rounding operation to chromosomes; second step is to count the location of duplicate gene (i.e., gene appears second or subsequent times); third step is to count the location of illegal gene whose value is larger than maximum dimension or less than 1; final step is to select gene that does not appear in chromosomes, and randomly place them into the gene location counted in second step for ensuring the chromosome legitimate after repair operator. The elite reserved strategy is used in the children. Ultimately new populations of children generation are formed for the next iterative genetic operation. Evolution process is executed many generations until the convergence condition is reached.

4. **Combination of Interactive Feature Extraction Using Quadratic Map and Feature Selection Using Genetic Algorithm.** Our algorithm for traditional GA has three main amendments: chromosome design, fitness function, and new repair operator. Figure 1 shows a system structure diagram of interactive feature extraction using quadratic map and interactive feature selection using genetic algorithm.

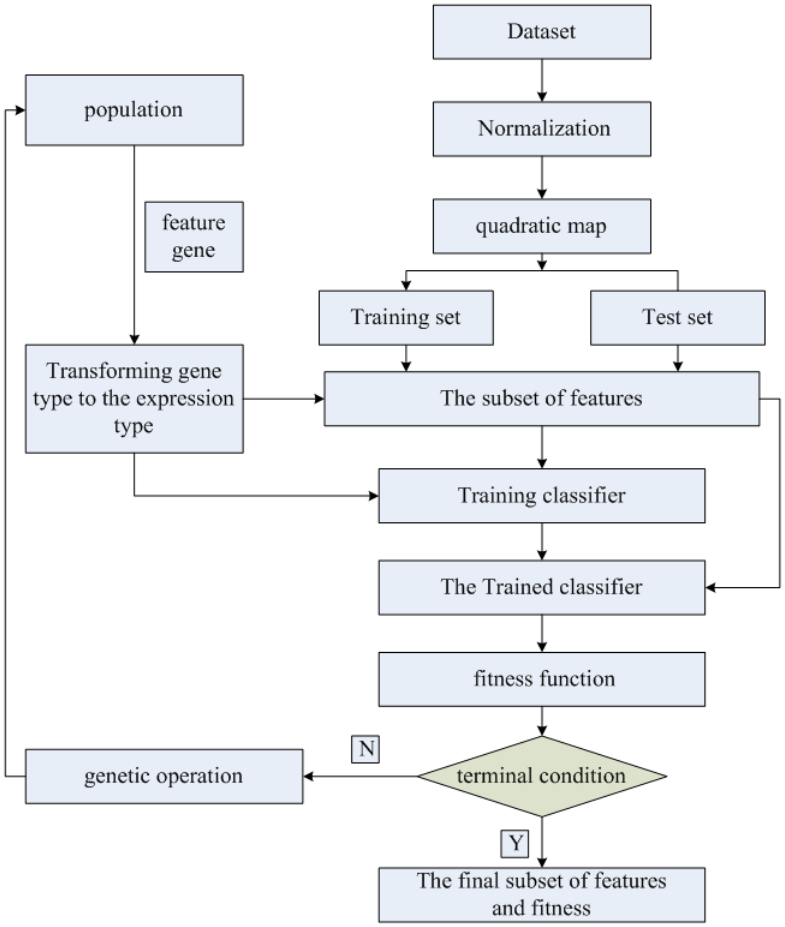

FIGURE 1. The diagram of the combination of interactive feature extraction using quadratic map and feature selection using genetic algorithm

The main steps of the proposed method are summarized as follows.

(1) Data normalization. The main advantage of normalization is to avoid large scale attributes affecting smaller scale ones, and avoid numerically calculating difficulty. Generally, each feature is linear normalized to $[0, 1]$.

(2) Quadratic map: According to the visual feature extraction Formula (1), we will get the new data with $d(d+3)/2$ dimensions by quadratic map for d dimensions data.

(3) Generating initial population: each chromosome denotes a set of d dimensions feature. The d dimensions are chosen from $d(d+3)/2$ dimensions.

(4) Fitness function evaluation of initial populations: Based on the selected features of the chromosomes of initial populations, the new data are divided into the training sample and testing sample sets. Training sets are used for training classifier, and testing sets are used for calculating the classification accuracy. Classification accuracy is fitness function of chromosomes.

(5) Genetic manipulation: This step searches better chromosomes by selection, crossover, mutation, and repair, replacement parent generation. The better solution and better feature are selected.

(6) Fitness function evaluation of genetic populations: Based on the selected features of the chromosomes of genetic populations, the new data are divided into the training sample and testing sample sets. Training sets are used for training classifier, and testing sets are used for calculating the classification accuracy.

(7) When the termination conditions such as the maximum number of iterations meet, the process ends. Otherwise, the process returns 5) genetic manipulation.

5. **The Experiment Results and Discussion.** In order to evaluate the proposed method, we use data sets from UCI [11], which are often used to perform comparison of various kinds of pattern recognition task in the literature. Table 1 summarizes the sample numbers, the feature and the categories of these data sets.

The normalization method uses $[0, 1]$ linear normalization, and we use 10 CV to evaluate the error rate of classifier. As a comparison, we test the classification performance of interactive barycentre features under the original feature ordering, and also test the classification performance of traditional feature extraction method as FLDA and GDA. The dimensions produced by all the feature extraction methods are the original feature dimensions, such as d. The radial basis kernel function with kernel parameters 4 is chosen for GDA method. For KNN we use nearest neighbor (1NN) and 3neighbour (3NN). For SVM we use radial basis function kernel in which kernel parameter is 0.5, hyper-parameter C is 100. LDA and KNN classifiers use PRTOOLS toolbox. SVM uses LIBSVM toolbox. The parameters setup of proposed method is shown below: a population has 20 chromosomes, and the maximum of iterations is 50, and selection operation is roulette, and crossover operation is a single point of intersection with crossover rate 1, and mutation operation is uniform variation with the mutation rate 0.02.

The experimental results of average error rate of 10 times 10 CV classification of these data sets are shown in Table 2 to Table 4. We can see from the table, the classification performance of our method is better than interactive features of star plot with original order, and also better than FLDA and GDA feature extraction methods. The classification performance of the interactive features is poor. As shown in Table 5, we compare the best classification performance reported in [5,12] with our methods. The reference [12] reported classification performance of almost all common classifiers on some data. [5] shows the classification results which used the dissimilarity representing method. The results show that our method is quite good.

Table 1. Data set information

Data set	Sample	Feature	Class
Wine	178	13	3
Breast-cancer-Wisconsin	683	9	2
Pima Indians diabetes	768	8	2

Table 2. The average error rate of Wine dataset under four kinds of features and four classifiers

Features	LDA	1NN	3NN	SVM
FLDA	0.0219	0.1758	0.1955	0.0337
GDA	0.0045	0.0062	0.0056	0.0056
Interactive features	0.0112	0.0618	0.0618	0.0393
Our method	0.0042	0.0054	0.0090	0.0000

Table 3. The average error rate of Breast-cancer-Wisconsin dataset under four kinds of features and four classifiers

Features	LDA	1NN	3NN	SVM
FLDA	0.0307	0.0565	0.0577	0.1362
GDA	0.0250	0.0366	0.0321	0.0293
Interactive features	0.0395	0.0439	0.0264	0.0293
Our method	0.0322	0.0205	0.0161	0.0249

TABLE 4. The average error rate of Pima Indians diabetes dataset under four kinds of features and four classifiers

Features	LDA	1NN	3NN	SVM
FLDA	0.2303	0.3096	0.2721	0.2643
GDA	0.2238	0.3069	0.2775	0.2266
Interactive features	0.2266	0.3333	0.2865	0.2301
Our method	0.2070	0.2617	0.2122	0.2174

TABLE 5. Comparisons of error rate with the results in the existing references

Data set	Our method	Reference [5]	Reference [12]
Wine	0.0000	0.000	0.011
Breast-cancer-Wisconsin	0.0161	0.002	0.025
Pima Indians diabetes	0.2070	0.211	0.224

6. **Conclusion.** The combination of interactive feature extraction using quadratic map and feature selection using genetic algorithm is proposed, based on the data visualization method of graphical representation of multivariate data. The experimental results of three kinds of UCI data sets show that the best classification error rate of the discriminating interactive barycentre feature reached 0.0%, 1.6% and 20.7%, respectively. It is better than the traditional feature extraction methods. In the future, the work includes the new interactive feature extraction method such as orthocenter, other feature selection methods such as particle swarm optimization or differential evolution, and the representation of geometric algebra instead of quadratic map.

Acknowledgment. This work was supported by National Natural Science Foundation of China (61273019, 61473339), China Postdoctoral Science Foundation (2014M561202), Hebei Province Postdoctoral Special Foundation (B2014010005), Hebei Province Top Young Talents 2013-17.

REFERENCES

[1] A. Fernández, Á. Gómez, F. Lecumberry et al., Pattern recognition in Latin America in the big data era, *Pattern Recognition*, vol.48, no.4, pp.1185-1196, 2015.

[2] Y. A. Ghassabeh, F. Rudzicz and H. A. Moghaddam, Fast incremental LDA feature extraction, *Pattern Recognition*, vol.48, no.6, pp.1999-2012, 2015.

[3] D. J. Janvrin, R. L. Raschke and W. N. Dilla, Making sense of complex data using interactive data visualization, *Journal of Accounting Education*, vol.32, no.4, pp.31-48, 2014.

[4] M. Stafoggia, A. Lallo and D. Fusco, Spie charts, target plots, and radar plots for displaying comparative outcomes of health care, *Journal of Clinical Epidemiology*, vol.64, no.7, pp.770-778, 2011.

[5] E. PeRkalska, R. P. W. Duin and P. Paclik, Prototype selection for dissimilarity-based classifiers, *Pattern Recognition*, vol.39, no.2, pp.189-208, 2006.

[6] G. Liu, H. Seiler and A. Wen, Distinct memory traces for two visual features in the drosophila brain, *Nature*, vol.439, no.7076, pp.551-556, 2006.

[7] C.-G. Park. Generalized quadratic mappings in several variables, *Nonlinear Analysis: Theory, Methods & Applications*, vol.57, nos.5-6, pp.713-722, 2004.

[8] J. L. Wang, P. L. Zhao, S. C. H. Hoi et al., Online feature selection and its applications, *IEEE Trans. Knowledge and Data Engineering*, vol.26, no.3, pp.698-710, 2014.

[9] C.-L. Huang and C.-J. Wang, A GA-based feature selection and parameters optimization for support vector machines, *Expert Systems with Applications*, vol.31, no.2, pp.231-240, 2006.

[10] G. Chen and J. Chen, A novel wrapper method for feature selection and its applications, *Neurocomputing*, vol.159, no.1, pp.219-226, 2015.

[11] M. Lichman, *UCI Machine Learning Repository*, http://archive.ics.uci.edu/ml, Irvine, CA, University of California, 2013.

[12] *Datasets Used for Classification Comparison of Results*, http://www.is.umk.pl/projects/datasets.html, 2015.