Genes Expression Data Processing Algorithm Based on Hybrid Particle Swarm Optimization

Zhiping Miao, Yiduo Qi, Airong Qian  
Bone Metabolism Lab, Key Laboratory for Space Bioscience & Biotechnology, School of Life Sciences, Northwestern Polytechnical University, Xi’an 710072, China.

Feng Zhang*  
School of Information Engineering, Yulin University, Yulin 719000, Shaanxi, China  
*Corresponding author(E-mail: tfnew21@sina.com)

Abstract
In order to improve the accuracy of structural and functional information of gene experiment data, and to better identify gene data related to function, a PSO-based hybrid algorithm (HPSO) was proposed and HPSO algorithm was applied to cluster analysis of gene expression data. And the problem of clustering analysis of gene expression data is reduced to an optimization problem, then a gene expression data processing algorithm based on HPSO is proposed, and an algorithm of gene expression data processing based on HPSO is proposed, and the HPSO algorithm was used to analyze the protein gene structure data, and a lot of useful rules were obtained. The experimental results show that the HPSO algorithm can cluster well, which opens up a new method for the analysis of gene expression data.

Key words: Data Mining, Gene Expression Data, Particle Swarm Optimization.

1. INTRODUCTION
Along with the development of genome and proteome research, and the rapid development of modern biotechnology, the high throughput technology has produced massive biological data, which provides a data base for uncovering the mystery of life (Chee and Yang, 1996). Biological data is rich in variety, high-throughput, high dimension and heterogeneous variability, which is far beyond the traditional analytical methods. The analysis of biological data becomes the bottleneck of current biological research, and its processing, mining and analysis and understanding of the increasingly urgent requirements. At present, there are some problems in the analysis of biological data (He and Stoevesandt, 2008). For example, the algorithm model used in data analysis is becoming more and more complicated, and the results obtained by the black box algorithm used in data analysis are difficult to make biological interpretation. The basic purpose of bioinformatics research is to use biological data to explain the phenomenon of life, and to explore the law of life.

Gene expression data refers to data obtained by directly or indirectly measuring the abundance of transcript mRNA in a cell by DNA microarray experiments (Uggar and Bittner, 1999). Gene expression data contain a wealth of information about gene activity, use of this information, you can observe the changes in the gene, the correlation between the analysis of genes to understand how under different conditions of gene activity are affected (Matsumura and Saithoh, 2005); In addition, gene expression data can reflect the current physiological state of cells. So the gene expression data can help researchers to understand the nature of many biological processes. How to extract the information about the structure and function of genes from the experimental data, how to identify the functionally related genes and how to get the exact expression of the physiological and biochemical pathways in the organism is a core problem in the field of gene chip technology research. Because of the large number of genes and the complexity of biological networks, cluster analysis technology has become a popular and effective technique for gene expression data processing as an exploratory data analysis method. At present, a large number of algorithms for gene expression data clustering are proposed in the field of bioinformatics. These algorithms are divided into three categories: gene-based clustering, sampled-based clustering, and biclustering (Mulligan and Mitsiades, 2007).

Through the study and analysis of massive gene expression data, gene expression and gene function can be linked to the unknown function of the new genes are classified into the known functions of the gene classification, at the same time, you can also find a new type of co-regulatory gene, So as to obtain a relatively complete gene regulatory network (Jun Sun and Wei Chen, 2012). At present, the research and analysis of gene expression data has become the core of biology, bioinformatics and system biology research. However, because of the few practical analysis tools for gene expression data, how to analyze and analyze the gene expression data is still faced with Great challenge. Because the genetic algorithm is applied to the clustering analysis of gene expression...
data, there are many unsatisfactory places, the researchers take into account another kind of intelligent Particle Swarm Optimization (PSO) algorithm which is more suitable for practical application (Moyes and Thelma Safadi, 2012). Compared with genetic algorithm, Particle Swarm Optimization algorithm does not need the operation of genetic operators, the complexity is relatively low, and the parameter setting is more flexible. But the PSO algorithm tends to premature convergence, and when the data is large, the efficiency of the algorithm will be greatly reduced. In this paper, we will discuss and study these problems, and improve the algorithm to obtain better clustering results.

2. EXPRESS DATA PRETREATMENT PROCESS

The expression data obtained after the DNA hybridization experiment are obtained by subtracting the background value from the foreground value of each point(Wang and Zineddin,2013), so that there will be a negative value or a small value, which is usually not biologically meaningful and should be removed. For these data points, they can usually be set to be zero or give them a unified value. For example, for oligonucleotide chip data, all data below 100 can be set to 100(Reis-Filho and Pusztai, 2011). The result of the microarray experiment was determined by measuring the signal intensity between the sample and the control sample. Assuming that the signal strength of the measured sample is \( T_1, T_2, \ldots, T_p \), the signal strength obtained from the reference sample is \( N_1, N_2, \ldots, N_p \), then the data conversion method is as follows:

(1) Logarithmic transformation: This transformation is the most common method for gene expression data transformation. First, the ratio of the signal intensity between the measurement sample and the control sample is calculated, \( \text{Ratio}_i = T_i / N_i \), we need the experimental results \( \bar{T} = \log \text{Ratio}_i \).

(2) Zero-value conversion: \( \bar{N} = \text{mean}(N_1, N_2, \ldots, N_p) \), the resulting data \( \bar{T} = T_i - \bar{N} \).

(3) DSGA conversion: The transformation is used for the identification of disease genes. The transformation of the original data is divided into two parts: the normal expression of the part, the degree of disease. Subsequent processing is performed on the degree of disease progression.

In the experiment, the expression of some points and other points there is a big difference, which is often caused by noise. In the traditional gene expression data processing is usually according to the Radio value to determine whether the point for the exception. When there are only a few particularly large Ratio values in a gene spectrum, they are outliers. In the subsequent development, new outlier identification methods have been introduced into the pretreatment of gene expression data. Commonly used methods include statistical methods, distance-based methods and density-based methods.

The ratio is used to represent the results of the DNA microarray experiment, which is the ratio of the signal strength of the sample to the control sample. In order to make the Ratio value about 0 symmetric, we have to deal with it, making less than 1 of the data becomes larger, instead, more than 1 of the smaller (Weight and Baehner, 2010). In order to reduce the complexity of data operation, we choose 2 as the base of the logarithm. When gene expression data are analyzed from time series, the result corresponds to the expression level at time zero.

(Vazquez and Carlos, 2015) proposed a self-learning hybrid model method for anomaly detection. The proposed method divides the data set \( D \) into two subsets of the normal set \( M \) and the exception set \( A \). Let it be generated by the distribution model \( M \), \( A \) respectively. The probability that any of the sample \( x_i \) of \( D \) belongs to \( A \) and \( M \) is respectively \( \lambda \) and \( (1-\lambda) \). then the distribution model \( D \) of \( D \) can be expressed as \( M, A \) mixed model:

\[
D = (1-\lambda)M + \lambda A
\]  

(1)

Firstly, this anomaly detection method divides the data set \( D \) into \( M_0(A_0) \), where \( M_0 = D, A_0 = \phi \), Then the anomaly samples are detected step by step from the \( M \) sets and classified into the anomaly set \( A \). Let \( t \) be the step \( D \) is divided \( M_0(A_0) \). Taking \( M \) and \( A \) as learning samples, the distribution function \( p_{M_i}(x) \) of \( A \) and \( M \) is obtained by using the suitable machine learning technology, and then find the distribution function \( p_{A_i}(x) \), then the probability \( L_t(D) \) of model \( D \) to produce data set \( D \) is:

\[
L_t(D) = \prod_{i=1}^{k} P_{d}(x_i) = (1-\lambda)^{\left| A \right|} \prod_{i=0}^{k} P_{M_i}(x_i) \left( \lambda^{\left| A \right|} \prod_{i=k+1}^{k} P_{A}(x_i) \right)
\]  

(2)

Where, \( \left| A \right|, \left| A \right| \) are the number of elements in sets \( D \) and \( F \), respectively. Since the initial exception set \( A_0 \) is an empty set, and \( P_{A_0} \) can not be obtained by machine learning, it needs to be specified in advance.
3. RESEARCH ON CLUSTERING ALGORITHM FOR GENE EXPRESSION DATA

In biology, the function of known genes is relatively small, to infer unknown knowledge, we must take full advantage of these relatively little known knowledge. Classification is the training of less data to come to the model of the entire data set, but for too large a gene expression data set, it is clear that the classification method does not apply, because the training model does not necessarily represent the entire data Space distribution. And cluster analysis is not required a priori knowledge, it directly through a certain law, the use of recursive partitioning method to divide the class, the target data with similar characteristics are classified into relatively homogeneous classes, that is, genes with similar expression properties are clustered together, this feature makes cluster analysis very suitable for simple and effective analysis of gene function. Gene expression data clustering analysis of the main purpose of the following two points: ① found some unknown cell state; ② detection of a class of functionally similar genes. How to identify the co-expression pattern of genes is the most critical problem in gene expression data clustering analysis. According to the common expression patterns, genes can be divided into different categories, which help us a deeper understanding of biological function and its relevance.

At present, a large number of clustering algorithms have been applied to the analysis of gene expression data. According to the nature of the algorithm, it can be divided into five types: partitioning method, hierarchical method, model-based method, density-based method, and a grid-based method. The most typical partition-based clustering algorithm is the K-means algorithm (He and Xu, 2011). K-means algorithm with the overall squared error as the evaluation criteria to optimize, the objective function is shown in Equation 3:

\[ E = \sum_{i=1}^{p} \sum_{o \in p_i} |o - \mu_i|^2 \] (3)

Where, \( E \) represents the sum of squared errors for all data objects, \( o \) is a data object in class \( p_i \), \( \mu_i \) is the cluster center of \( p_i \). With \( E \) as the objective function, the objective function is minimized by iteration, even if interclass differences are minimized.

Tracking Kalman filter for linear systems, and are assuming a Gaussian posterior probability distribution at any time can be determined by the mean and variance of the distribution. Kalman filter algorithm that is based on the following assumptions:

1) Target transfer noise from the \( v_{t-1} \) and measurement noise \( n_t \) is Gauss distribution with known parameters.
2) The state transition equation \( f_t(x_{t-1}, v_{t-1}) \) is a linear function of \( x_{t-1} \) and \( v_{t-1} \)
3) The measurement equation \( h_t(x_t, n_t) \) is a linear function of \( x_t \) and \( n_t \)

Then,

\[ x_t = F x_{t-1} + v_{t-1} \] (4)
\[ z_t = H x_t + n_t \] (5)

Where, Among them, \( F \) and \( H \) are matrix definition of linear function, \( v_{t-1} \) and \( n_t \) are independent of each other, and the mean value is 0, \( Q \) and \( R \) are variance of a Gaussian distribution. Then it is according to Bayesian estimation theory, we can get a Kalman filter algorithm:

(1) State Prediction
a) Target state and variance forecast

\[ x_{t|t-1} = F x_{t-1|t-1} \] (6)
\[ P_{t|t-1}^x = Q + FP_{t-1|t-1}F^T \] (7)

b) Predict the measured value and variance

\[ z_{t|t-1} = H x_{t|t-1} \] (8)
\[ P_{t|t-1}^z = R + FP_{t|t-1}^x H^T \] (9)

c) Correlation matrix is a priori predictions

\[ P_{t|t-1}^{xz} = P_{t|t-1}^x H^T \] (10)

(2) measurement update
a) Calculating the Kalman gain

\[ K_t = P_{t|t-1}^z [P_{t|t-1}^z]^{-1} = P_{t|t-1}^x H^T [R + HP_{t|t-1}^z P_{t|t-1}^{xz} H^T]^{-1} \] (11)
b) Using of time $t$ measured values, update the target state prediction and estimation variance

$$x_{t|t-1} = x_{t-1} + K_t(z_t - z_{t-1})$$  \hspace{1cm} (12)

$$P_{t|t-1} = P_{t-1} - K_tP_{t|t-1}K_t^T = (1-K_tH_t)P_{t|t-1}$$  \hspace{1cm} (13)

Since the Kalman filter algorithm is simple, small amount of calculation has been widely used in target tracking. Kalman filter algorithm is a Bayesian filter theory of linear, Gaussian distribution analytical results obtained under the assumption, therefore, is the best estimate of the Kalman filter algorithm for linear Gaussian filter under the concept.

The algorithm is composed of the following steps (Monson and Seppi, 2004):

1) Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2) Assign each object to the group that has the closest centroid.
3) When all objects have been assigned, recalculate the positions of the K centroids.
4) Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. The k-means algorithm can be run multiple times to reduce this effect.

4. CLUSTERING OF GENE EXPRESSION DATA BASED ON PARTICLE SWARM OPTIMIZATION

Particle filter is a method of Monte Carlo simulation based on recursive Bayesian filtering, the key idea is to use a set of weighted sum of the weights associated with a random sample to represent posterior probabilities. When the sample size is very large, this probability is estimated to be equivalent to the posterior probability density. Can assume a state independent from the posterior probability distribution $\rho(X_i \mid Z_i)$ of the N independent random sample of $\{X_i^{(n)}\}, i=1,2,...,N^i$, the state probability density distribution can be approximated as:

$$\rho(X_i \mid Z_i) \approx \frac{1}{N} \sum_{i=1}^{N} \delta(X_i)$$  \hspace{1cm} (14)

Where $\delta(\cdot)$ is the Dirac impulse function, the corresponding function is expected:

$$\tilde{T}_{\pi(x)} = \int g_i(X_i) \tilde{p}(X_i \mid Z_i) dX_i = \frac{1}{N} \sum_{i=1}^{N} g_i(X_i^{(n)})$$

By the law of large numbers can guarantee convergence, the convergence does not depend on the state dimension, can be easily applied to higher dimensional case.

1) The importance sampling

Because of the need to estimate the probability distribution of a sample is often very difficult or even impossible, to avoid the following direct importance sampling method of sampling difficulties, from another random sampling is easier to extract the distribution. If the density of the sample $\rho(X_i \mid Z_i)$ from the posterior probability is difficult to obtain a sample directly from the particles, where the probability of an easier introduction of the sample and from the sample distribution of $q(x_i \mid Z_i)$, $q(x_i \mid Z_i)$ is called here the importance of the distribution. In this case the above equation becomes:

$$I(g_i) = \int g_i(X_i) \frac{p(X_i \mid Z_i)}{q(X_i \mid Z_i)} q(x_i \mid Z_i) dx_i$$  \hspace{1cm} (15)

One of the important weights:

$$\omega'(X_i) = \omega'(X_i^{(n)}) = \frac{p(X_i \mid Z_i)}{q(X_i \mid Z_i)} = \frac{p(X_i \mid Z_i)}{\rho(Z_i)} \times \frac{1}{q(X_i \mid Z_i)} = \frac{1}{\rho(Z_i)} \times \frac{p(X_i \mid Z_i)}{q(X_i \mid Z_i)}$$  \hspace{1cm} (16)

According to Bayesian theory, normalization constants $\rho(Z_i)$ denominator in the formula can be expressed as:

327
\[
\rho(Z_t) = \int \rho(Z_t \mid X_t) \rho(X_t) dX_t = \int \omega_k(X_t) q(X_t \mid Z_t) dX_t E_{q(z_t)} \left[ \omega(x_t) \right]
\]  

(17)

Take the proposal from a group of independent and identically distributed \(q(X_t \mid Z_t)\) distributed particle swarm \(\{X^{(i)}_t, i = 1, 2, ..., N\}\), is estimated as follows:

\[
I_k(g_t) = \frac{1}{N} \sum_{i=1}^{N} g_k(x^{(i)}_t) = \sum_{i=1}^{N} g_k \left( X^{(i)}_t \right) \sigma^{(i)}_t
\]

(18)

Wireless sensor networks have a localized target tracking features, namely multi-target intersection outside the area, you can multi-target tracking problem, seen as separate multiple single target tracking "So, to solve the wireless sensor networks multi-target tracking problem the core is how to measure the value of the acoustic sensor node to make the appropriate separation.

Consider the following case, the target at the time \(T_1\) and \(T_2\) into the target intersection area, when the node \(i\) is located in the intersection area, while the acoustic signal obtained from the \(T_1\) and \(T_2\), and outputs the measured value \(z_i(t)\).

Localized by the target tracking shows that when \(t < T_1\), \(T_1\) and \(T_2\) of the track, a single target tracking, target tracking at this time can be achieved, that the amount of \(T_1\) time before the sensor node to obtain the measured value has been obtained at the \(T_2 = t\) time the measured value maneuvering targets, and thus get the location of the target time, speed and other information.

The above ideas with particle filter tracking algorithms, particle filter tracking algorithm is based on the prediction of multi-objective, as follows:

For each target \(K\), perform the following algorithm

1) \(t = 0\), set reference probability density function of the sample \(q(x_t \mid x_{t-1}, z_t) = \rho(x_t \mid x_{t-1})\) and initialize particle distribution

\[
x_0^i \rho \left( x_0^i \right), w_0^i = 1/N, i = 1, ..., N
\]

(19)

2) when \(t > 0\),

\[
\text{for } i = 1 : N \times N, q \left( x^{(i)}_t, z_t \right)
\]

Calculating weights

\[
w^{(i)}_{k,t+1} = w^{(i)}_{k,t} \frac{p \left( z_t \mid x^{(i)}_{k,t} \right) p \left( x^{(i)}_{k,t} \mid x^{(i)}_{k,t-1} \right)}{q \left( x^{(i)}_{k,t} \mid x^{(i)}_{k,t-1}, z_t \right)}
\]

(20)

3) according to the type of normalized weights

\[
\bar{w}^{(i)}_{k,t} = \frac{w^{(i)}_{k,t}}{\sum_{i=1}^{N} w^{(i)}_{k,t}}
\]

(21)

4) predict the measured value

According to the target maneuvering model, predict the target position

\[
\hat{x}_{k,t+1} = f \left( \hat{x}_{k,t} \right) + w_{k,t}
\]

(22)

According to the measurement model, measurement of nodes in \(s_j\) this forecast the amount of value \(k\)

\[
z_{k,j+1} = h \left( x_{k,t}, s_j \right)
\]

(23)

5) repeat steps 2-4, until the end of track.

From the above analysis, it can be seen that the PSO algorithm has the memory, and the genetic algorithm can change the individual values of the original group with the continuous variation of the population. Although the particle swarm of multi-objective fitness calculation and optimization has higher computational efficiency, but in the latter part of the iterative process will fall into local optimum, is not conducive to solving multi-objective optimization of the global optimal solution. And genetic algorithm with crossover and mutation operation,
eliminating the iterative process fitness solution does not adapt, not easy to fall into local optimum, can achieve the global optimal solution.

With the continuous replacement of various technologies, the research of various algorithms is constantly deepened, domestic and foreign scholar's study of multi-objective optimization algorithm is focuses mainly the basis of theoretical analysis, this algorithm can be summarized as the design of the algorithm is improved and design of the new algorithm. For the characteristics of GA and PSO, in this paper, crossover and mutation operator and population segmentation strategy are introduced into PSO algorithm, two kinds of algorithms are mixed to solve the multi-objective optimization problem, a PSO-based hybrid algorithm(HPSO) is proposed. HPSO algorithm flow is shown in Figure 1.

![HPSO Algorithm Flowchart](image)

**Figure 1. HPSO algorithm flowchart**

5. SIMULATION AND PERFORMANCE ANALYSIS

Let the gene expression data set be \( \{g_1, g_2, ..., g_n\} \), is an \( n \times m \) matrix, the expression pattern of \( N \) gene was expressed in \( M \) experimental samples. The \( i \)-th dimensional attribute value of the gene \( g_i \) represents the expression level of the gene \( g_i \) under the \( j \)-th experimental sample, Using \( w_{i,j} \) representation,where, \( i = 1, 2, j = 1, 2, ..., m \). Suppose that this data set is partitioned into \( K \) clusters,where the first \( k \) classes with \( Q_k \) representation, then the clustering center vector of the class \( Q_k \) is represented as \( \mu_{g_{gen}} = (q_{gen_{k,1}}, q_{gen_{k,2}}, ..., q_{gen_{k,n}}) \), the number of individuals contained in \( Q_k \) is represented by \( T_k \), Specific values for each dimension such as the formula 24:

\[
q_{gen_{k,i}} = \sum_{i=1}^{n} q_{i,j} w_{i,j} / T_k
\]  

(24)
We use Euclidean distance as the similarity measure for clustering algorithm of gene expression data based on PSO, the sum of the distance between each individual and each cluster center in each class is called intraclass difference, The class difference within the class k can be expressed by the formula 25:

\[ CY_k = \sum g_{ik} \sum_{j=1}^{n_{ik}} (w_{ij} - q_{gen_{ik}})^2 \]  

(25)

After the current gene expression data set is divided, The intra-class variance of the k classes is called Total Within-Cluster Variation (TWCV). Then the overall intraclass difference can be calculated using Equation 26:

\[ TCY_k = \sum_{k=1}^{t} CY_k = \sum_{k=1}^{t} \sum_{i=1}^{G_{i}} \sum_{j=1}^{n_{ij}} (w_{ij} - q_{gen_{i}})^2 \]  

(26)

We use Equation 26 as the objective function of the PSO-based gene expression data clustering algorithm. The evolution of the algorithm is to optimize the formula. The smaller the objective function value is, the smaller the total intraclass difference value is, the better the clustering effect is.

In order to verify the effectiveness of the HPSO gene expression data based on clustering algorithm, the improvement of the performance of HPSO algorithm is better, we selected 4 gene expression data sets of different algorithms for clustering simulation test. Test environment for MATLAB 11 software environment. The application of the algorithm is tested by K-means algorithm and HPSO algorithm. And some typical parameters of the experimental results were analyzed and compared, and the performance of different clustering algorithms in the analysis of gene expression data was evaluated. Based on the experience and experimental test case, the relevant parameters are set as follows: The number of particles \( n = 100 \), inertia weight \( \omega = 0.92 \). In this paper, four datasets are selected, one synthetic data set, and three natural data sets. The properties of each data set are shown in Table 1. Figure 2 is the convergence curve of two algorithms when clustering the dataset, where the number of clusters is the number of external standard clusters.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Number of data (n)</th>
<th>Data dimension (m)</th>
<th>Standard class number (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1024</td>
<td>256</td>
<td>16</td>
</tr>
<tr>
<td>B</td>
<td>5126</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>30109</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>600</td>
<td>16</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1. Dataset properties

![Figure 2. Convergence curves of data sets](image)

It can be seen from the convergence curves of each algorithm on different data sets that the convergence speed of K-means algorithm is very fast, and the algorithm tends to converge after several iterations, but the clustering result is not good, but converges to A local optimum point, the HPSO algorithm has obvious advantages in terms of convergence speed and clustering result. The convergence speed is close to K-means algorithm, and the clustering results are very good.

6. CONCLUSIONS

Clustering algorithm is a kind of data mining method to classify unsupervised data according to its similarity. It can classify similar data into one kind. When the gene expression data is processed by clustering algorithm, it can cluster the genes with the similar expression pattern into one class, so as to further infer the function of the gene and discover the interrelationship between the genes. Based on the deep analysis of the basic theory of PSO, PSO algorithm is applied to cluster analysis of gene expression data, and PSO algorithm is...
improved. In order to avoid the premature convergence of the algorithm, a hybrid PSO algorithm (HPSO) is proposed, to help trapped particles jump out of local optimum, combining the improved algorithm and K-means algorithm, proposes a clustering algorithm of HPSO gene expression data based on the advantages of the proposed algorithm can effectively K-means algorithm and PSO algorithm. And selects the different data sets, to K-means and HPSO cluster simulation test, has carried on the analysis to the experiment result, verified the validity and the superiority of the HPSO algorithm.

ACKNOWLEDGEMENTS

This work is partially supported by Funding Project for Department of Education of Shaanxi Province of China (#15JK1861), the National Natural Science Foundation of China (Grant No.U1501253, No.51651901). Thanks for the help.

REFERENCES


Wa

