# AN IMPROVED QUANTUM-INSPIRED EVOLUTIONARY ALGORITHM FOR CLUSTERING GENE EXPRESSION DATA

W.G. Zhou, C.G. Zhou, G.X. Liu, H.Y. Lv and Y.C. Liang
*College of Computer Science and Technology, Jilin University, Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Changchun 130012, P. R. China*

**Abstract** Microarray technologies have made it straightforward to monitor simultaneously the expression pattern of thousands of genes. So an important task is to cluster gene expression data to identify groups of genes with similar patterns and hence similar functions. In this paper, an improved quantum-inspired evolutionary algorithm (IQEA) is first proposed for minimum sum-of-squares clustering. We have suggested a new representation form and added an additional mutation operation in IQEA. Experiment results show that IQEA appears to be much more robust in finding optimum or best-known solutions and be superior to conventional k-means and self-organizing maps clustering algorithms even with a small population.

**Keywords:** gene expression, clustering, quantum-inspired evolutionary algorithm.

## 1. INTRODUCTION

In the field of bioinformatics, clustering algorithms have received renewed attention due to the breakthrough of microarrays data. Microarrays experiments allow for the simultaneous monitoring of the expression patterns of thousands of genes. Since a huge amount of data is produced during microarray experiments, clustering methods are essential in the analysis of gene expression data. The goal is to extract the fundamental patterns inherent in the data and to partition the elements into subsets referred to as clusters. In gene expression, elements are usually genes. The vector of each gene contains its expression levels under each of the monitored conditions. Several clustering algorithms have been proposed for gene expression data analysis, such as hierarchical



*G. R. Liu et al. (eds.), Computational Methods,* 1351–1356.
© 2006 *Springer. Printed in the Netherlands.*

clustering, self-organizing maps, k-means, and some graph theoretic approaches. In this paper, we propose an improved quantum-inspired evolutionary algorithm (IQEA) for clustering gene expression data. The IQEA has been shown to perform well and be superior to k-means and self-organizing maps clustering algorithms even with a small population.

## 2. MINIMUM SUM-OF-SQUARES CLUSTERING

Clustering can be considered as a combinatorial optimization problem, in which an assignment of data vectors to clusters is desired, such that the sum of distance square of the vectors to their cluster mean (centroid) is minimal. Let $P_k$ denote the set of all partitions of X with $\mathrm{X} = \{x_1, \ldots, x_n\}$ denoting the data set of vectors with $x_i \in R^m$ and $C_i$ denoting the $i$th cluster with mean $x_i$. Then we can formulate the problem as the search for an assignment $p$ of the vectors to the clusters with $C_i = j \in \{1, \ldots, n\} | p[j] = i$. Thus the objective function becomes:

$$\min_p \sum d^2(x_i, \hat{x}_{p[i]}). \tag{1}$$

This combinatorial optimization problem is called the minimum sum-of-squares clustering (MSSC) problem as shown in the work by Merz [1] and has been proven to be NP-hard. Since the novel quantum-inspired evolutionary algorithm has been shown to be effective as compared to the conventional genetic algorithm, the application of QEA to the MSSC appears to be promising.

## 3. QUANTUM-INSPIRED EVOLUTIONARY ALGORITHM

Quantum-inspired evolutionary algorithm (QEA) is based on the concept and principles of quantum computing such as a quantum bit and superposition of states. Like other evolutionary algorithms, QEA is also characterized by the representation of the individual, the evaluation function and the population dynamics. A Q-bit is defined as the smallest unit of information in QEA, which is defined with a pair of numbers $(\alpha, \beta)$. It may be in the '1' state, in the '0' state, or in any superposition of the two. A Q-bit individual is a string of Q-bits. The state of a Q-bit can be changed by the operation with a quantum gate, such as NOT gate, Rotation gate, and hadamard gate, etc. Rotation gate is often used to update the Q-bit as follows:

$$\begin{bmatrix} \alpha_i' \\ \beta_i' \end{bmatrix} = \begin{bmatrix} \cos(\Delta\theta_i) & -\sin(\Delta\theta_i) \\ \sin(\Delta\theta_i) & \cos(\Delta\theta_i) \end{bmatrix} \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}. \tag{2}$$

*Table 1.* The angle parameters used for rotation gate.

| $x_i$ | $b_i$ | f(x) ≥ f(b) | $\Delta\theta_i$ | $x_i$ | $b_i$ | f(x) ≥ f(b) | $\Delta\theta_i$ |
|---|---|---|---|---|---|---|---|
| 0 | 0 | false | $\theta 1$ | 1 | 0 | false | $\theta 5$ |
| 0 | 0 | true | $\theta 2$ | 1 | 0 | true | $\theta 6$ |
| 0 | 1 | false | $\theta 3$ | 1 | 1 | false | $\theta 7$ |
| 0 | 1 | true | $\theta 4$ | 1 | 1 | true | $\theta 8$ |

The angle parameters used for rotation gate are shown in Table 1. Where $x_i$ and $b_i$ are the $i$th bit of the best solution b and the binary solution $x$ respectively. The structure of QEA is shown in previous work by Han [2].

## 4. THE GENE EXPRESSION DATA SETS

The first data set denoted as HL-60 is described in the work by Tomayo [3] and it contains data from macrophage differentiation experiments. It consists of 7229 genes and expression levels at four time points. We apply a variation filter which discarded all genes with an absolute change in maximum and minimum expression level less than 30. The number of genes which pass the filter is 2792. The vectors are then normalized to have mean 0 and variance 1. The second data set denoted as Yeast is described in the work by Cho [4]. It consists of 6602 yeast genes measured at 17 time points over two cell cycles. The 90-minute and 100-minute time points are excluded because of difficulties with scaling. Afterwards, we use a variation filter to discard all genes with an absolute expression level change less than 100, and an expression level of max/min $<$2.0. The number of genes that pass the filter is 2947. Again, the vectors are normalized to have mean 0 and variance 1.

## 5. IQEA FOR CLUSTERING GENE EXPRESSION DATA

In the experiments, we compare the improved IQEA with k-means and self-organizing maps (SOM) algorithms. IQEA and k-means algorithms are implemented in Matlab 6.5. The self-organizing maps algorithm is available in the software package Gene Cluster 2.0.

### 5.1 Initialization, Representation and Fitness Function

The k-means algorithm is first run ten times and so ten initial solutions are produced for each data set. Each solution is composed of $n$ mean vectors where

$n$ is the number of clusters. Given the mean vectors, the cluster memberships can be calculated by calculating the nearest distance of each gene vector with all mean vectors. The representation used in the IQEA is straightforward. There are $n^*10$ mean vectors in all for the ten initial solutions. Then we number each mean vector as 1, 2, 3 ... $n^*10$ and put them into the set $V$ where $V = (v_1, v_2, \ldots, v_{n^*10})$. So $n^*10$ Q-bits are used in each Q-bit individual. A Q-bit individual has the following form in the $t$th generation:

$$\begin{bmatrix} \alpha_1^t & \alpha_2^t & \cdots & \alpha_n^t & \alpha_{n+1}^t & \cdots & \alpha_{2\times n}^t & \cdots & \alpha_{9n+1}^t & \cdots & \alpha_{10\times n}^t \\ | & | & | & | & | & | & | & | & | & | \\ \beta_1^t & \beta_2^t & \cdots & \beta_n^t & \beta_{n+1}^t & \cdots & \beta_{2\times n}^t & \cdots & \beta_{9n+1}^t & \cdots & \beta_{10\times n}^t \end{bmatrix} \tag{3}$$

where $\alpha$ is initially given as a random number between 0 and 1 initially and $\beta$ could be computed according to Equation (2). The fitness function used in the IQEA is the MSSC error provided in Equation (1).

## 5.2 Make and Repair Operation

To obtain the binary string, the step of 'Make(x)' by observing the states of Q-bit can be implemented for each Q-bit individual as follows:

**If** (random (0, 1) $< |\beta_i|^2$ && $k < n$) **then** $x(i) \leftarrow 1; k \leftarrow k + 1;$
**else** $x(i) \leftarrow 0;$

where the variant $k$ is used to guarantee that the number of '1' in each binary string must be less than or equal to $n$. Afterwards, if the number of '1' is less than $n$ in some strings, an additional mutation operation 'Repair (x)' should be performed to be sure that there should be and only be $n$ '1' in each binary string as follows:

**while** k < n **do**
randpos $\leftarrow$ random (1, $n^*10$); **if** x (randpos) = 0 **then** x (randpos) $\leftarrow 1$; $k \leftarrow k + 1$;
**end**

## 5.3 Evaluated and Updated Operation

Then the evaluated operation is executed to calculate the fitness value according to Equation (1) for each Q-bit individual and so the best individual can be selected. The updated operation is used to update Q-bit states of each individual by rotation gate in Equation (3). The angle parameter in Table 1 is

*Table 2.* Experiment results for the two data sets.

| Dataset | Algorithm | Best obj | Avg obj | Excess |
|---|---|---|---|---|
| HL-60 | IQEA | 1514.3 | 1598.5 | 5.56% |
| | K-means | 1514.6 | 1604.3 | 5.92% |
| | SOM | 1523.2 | 1624.0 | 6.62% |
| Yeast | IQEA | 15741.0 | 15860.0 | 0.76% |
| | K-means | 15752.0 | 15905.0 | 0.97% |
| | SOM | 15801.0 | 16002.0 | 1.27% |

set as follows according to experiment observation:

$$\text{If } \alpha^*\beta > 0 \text{ then } \theta_3 = 0.01\ \pi; \theta_5 = 0.01\ \pi;$$
$$\text{If } \alpha^*\beta < 0 \text{ then } \theta_3 = 0.01\ \pi; \theta_5 = 0.01\ \pi;$$

## 6. EXPERIMENT RESULTS

In all experiments, IQEA is run with a population size of 20. The k-means and IQEA algorithms are all terminated when the maximum generation of 50 is arrived. The three algorithms are all run 10 times, and the best and the average objective value (fitness value) are produced. In Table 2, the results are displayed for the described two data sets. Excess value is the percentage of the average objective value above the best objective value. It is obvious that IQEA performs better than the other two algorithms.

## 7. CONCLUSIONS

In this paper, an improved quantum-inspired evolutionary algorithm (IQEA) by using a new representation form and adding an additional mutation operation is proposed for clustering gene expression data. We present experimental evidence that the proposed algorithm is effective and produces better solutions than the conventional k-means and self-organizing maps clustering algorithms even with a small population.

## REFERENCES

1. P. Merz (2003), Analysis of gene expression profiles: an application of memetic algorithms to the minimum Sum-of-squares clustering problem. *Biosystem*, 72, pp. 99–109.

2. K. Han and J. Kim (2002), Quantum-inspired evolutionary algorithm for a class of combinatorial problem. *IEEE Transactions on Evolutionary Computation*, 6, pp. 580–593.
3. P. Tamayo and D. Slonim (1999), Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *PNAS*, 96, pp. 2907–2912.
4. R.J. Cho, E.A. Winzeler and R.W. Davis (1998), A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2, pp. 65–73.