Artificial Intelligence in Medicine 73 (2016) 23–33

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

A comparative analysis of chaotic particle swarm optimizations for detecting single nucleotide polymorphism barcodes

Li-Yeh Chuang [a], Sin-Hua Moi [b], Yu-Da Lin [b,*], Cheng-Hong Yang [b,*]

[a] Department of Chemical Engineering & Institute of Biotechnology and Chemical Engineering, I-Shou University, No.1, Sec. 1, Syuecheng Rd., Dashu District, Kaohsiung City 84001, Taiwan
[b] Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, No.415, Jiangong Rd., Sanmin Dist., Kaohsiung City 80778, Taiwan

ARTICLE INFO

Article history:
Received 14 April 2016
Accepted 29 September 2016

Keywords:
Particle swarm optimization
Chaos
Single nucleotide polymorphism
Single nucleotide polymorphism barcode detection

ABSTRACT

Objective: Evolutionary algorithms could overcome the computational limitations for the statistical evaluation of large datasets for high-order single nucleotide polymorphism (SNP) barcodes. Previous studies have proposed several chaotic particle swarm optimization (CPSO) methods to detect SNP barcodes for disease analysis (e.g., for breast cancer and chronic diseases). This work evaluated additional chaotic maps combined with the particle swarm optimization (PSO) method to detect SNP barcodes using a high-dimensional dataset.

Methods and material: Nine chaotic maps were used to improve PSO method results and compared the searching ability amongst all CPSO methods. The XOR and ZZ disease models were used to compare all chaotic maps combined with PSO method. Efficacy evaluations of CPSO methods were based on statistical values from the chi-square test (χ^2).

Results: The results showed that chaotic maps could improve the searching ability of PSO method when population are trapped in the local optimum. The minor allele frequency (*MAF*) indicated that, amongst all CPSO methods, the numbers of SNPs, sample size, and the highest χ^2 value in all datasets were found in the Sinai chaotic map combined with PSO method. We used the simple linear regression results of the *gbest* values in all generations to compare the all methods. Sinai chaotic map combined with PSO method provided the highest β values ($\beta \geq 0.32$ in XOR disease model and $\beta \geq 0.04$ in ZZ disease model) and the significant *p*-value (*p*-value < 0.001 in both the XOR and ZZ disease models).

Conclusion: The Sinai chaotic map was found to effectively enhance the fitness values (χ^2) of PSO method, indicating that the Sinai chaotic map combined with PSO method is more effective at detecting potential SNP barcodes in both the XOR and ZZ disease models.

1. Introduction

Deoxyribonucleic acid (DNA) carries the inherited determining factors for specific personal characteristics. Genome-wide association studies (GWAS) are widely used to examine the genetic variants, known as single nucleotide polymorphisms (SNPs), in different individuals and the variant associations between genes and diseases [1]. Previous studies have shown that SNPs in human DNA sequences are significant impact factors for various diseases [2]. Case-control studies are widely used to evaluate differences between SNP expressions in model hypotheses of genetic effects [3].

Recent studies have focused on establishing associations between genetic loci in various diseases [4,5]. SNP barcodes were introduced as the combination of SNPs with genotypes to represent the association between genes [6]. SNP barcodes could suppress or increase the genetic effect of particular genes. However, many conventional analysis approaches show the difficultly of handling the huge computational requirements for the permutation testing of large datasets to evaluate SNP barcodes. Association analysis entails complex computations, especially for high-order associations using high-dimensional datasets.

Data mining and machine learning approaches have proposed various algorithms to facilitate engineering fields for solving com-

[*] Corresponding authors.
E-mail addresses: chuang@isu.edu.tw (L.-Y. Chuang), moi9009@gmail.com (S.-H. Moi), e0955767257@yahoo.com.tw (Y.-D. Lin), chyang@cc.kuas.edu.tw (C.-H. Yang).

http://dx.doi.org/10.1016/j.artmed.2016.09.002
0933-3657/© 2016 Published by Elsevier B.V.

putational problems. These algorithms are designed to efficiently identify patterns within the problems such as the travelling salesman problem [7], ordinal regression [8], ν-support vector regression [9], gene–gene interaction [10–13], ν-support vector classification [14], classification [15,16], image segmentation [17], Steganalysis of least significant bit [18,19], tRNA prediction [20], association rules in medicine [21], and so on. Efforts to improve algorithms mainly focus on enhancing efficiency, accuracy, and precision.

Particle swarm optimization (PSO) method facilitates the computation of high-order SNP barcode associations. Prior verification studies such as those related to facial emotion perception [22] and susceptibility to chronic disease [23] have demonstrated the effectiveness of PSO method to detect the SNPs barcode. Thus, such limitations on large statistical evaluations for high-order SNP barcodes can be overcome by the further development of efficacious evolutionary computing methods. PSO method can optimize a candidate solution through iterative attempts based on the robust mathematics of swarm intelligence. Recently, PSO method has been improved by several mathematical techniques including hybridization, along with combinatorial, multicriteria and constrained optimization [24]. Hybridization approaches mainly combines the desirable properties of several approaches to compensate for their individual weakness, i.e., the local optima problem. Most hybrid PSO methods provide better solutions for improving PSO methods in high-dimensional problems.

In our previous studies, the chaotic PSO (CPSO) method was found to effectively identify SNP barcodes in specific diseases with complex biological relationships [25–27]. In these works, the original PSO and two CPSO methods, including Logistic-PSO (logistic map) and DBM-PSO (double-bottom map), were used to compare the SNP barcode detection ability for breast cancer. CPSO method showed superior SNP barcode detection than PSO method, and DBM-PSO method provides better detection than Logistic-PSO method. These works indicated that original PSO method is easily trapped into local optima, and that chaotic maps combined with PSO method can prevent this problem. Moreover, based on hybridization theory, the CPSO method can use chaotic maps to improve on the original PSO method in terms of searching for a more significant SNP barcode. In this study, we used the nine chaotic maps to improve the original PSO method, and all CPSO methods are compared in terms of their efficiency for detecting significant SNP barcodes. The simulation datasets for XOR and ZZ diseases are generated to compare original the PSO and all CPSO methods. This study evaluates the effectiveness of additional chaotic maps combined with PSO method for the detection of SNP barcodes using high-dimensional datasets.

2. Methods

2.1. Problem definition

The non-deterministic polynomial-time hard (NP-hard) and feature selection problems are key issues in high-order SNP barcode identification [5,28]. Homozygous reference genotype ('AA'), heterozygous genotype ('Aa'), and homozygous variant genotype ('aa') are three common genotype classifications. The genotype at SNP_i is assigned as $G_i \in \{1, 2, 3 \mid 1 = \text{'AA'}, 2 = \text{'Aa'}, 3 = \text{'aa'}\}$, where i is i^{th} SNP in all SNPs. An SNP barcode is defined as a set $E = \{e_1, e_2, e_3, \ldots, e_m\}$, where $e_i = \{\text{SNP}_i, G_i\}$ and m indicates the m-dimensional SNP selection. The selection of m SNPs ($m \geq 2$) with the χ^2 for SNP barcode identification is used to evaluate the significance level of the association between SNPs under diseases. A fitness function $f(E)$(f: $\mathrm{R}^m \to \mathrm{R}$) is designed by the χ^2 function and the highest χ^2 value among all avaiable values of E is treated as the goal E^*.

```
Begin
Initialize PSO parameters and particles' vectors
while ( the stopping criterion is not met ){
    Evaluate particles' fitness
    for ( $n = 1$ ; $n \leq$ number of particles ; $n$++ ){
        Update pbest
        Update gbest
        for ($d = 1$; $d \leq$ particle's dimension ; $d$++){
            Update particle positions
        }
    }
    Update inertia weight value
}
End
```

Fig. 1. PSO pseudo-code.

2.2. Particle swarm optimization

PSO method is an evolutionary algorithm that iteratively attempts to optimize a candidate solution by sharing valuable information between individuals [29]. Thus, the shared information can lead individual particles toward better solutions. Particle P is a result in which the value of P is evaluated by the fitness function. The $pbest_i$ represents the previous best vector of i^{th} particle and is denoted as $pbest_i = (pbest_{i1}, pbest_{i2}, \ldots, pbest_{iD})$. The $gbest$ is defined from the best vector among all $pbest$ values and is denoted as $gbest = (gbest_1, gbest_2, \ldots, gbest_D)$. The current vector of the i^{th} particle is denoted as $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, where $x \in (X_{\min}, X_{\max})^D$, in which $X_{\min}$ and $X_{\max}$ are respectively the minimum and maximum SNP numbers, and D is the number of dimensions. The velocity of the i^{th} particle is represented as $v_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$, $v \in [V_{\min}, V_{\max}]^D$. The velocity and position of a particle can be updated by Eqs. (1) and (2).

$$v_{id}^{new} = w \times v_{id}^{old} + c_1 \times r_1 \times \left(pbest_{id} - x_{id}^{old}\right) + c_2 \times r_2 \times \left(gbest_d - x_{id}^{old}\right) \tag{1}$$

$$x_{id}^{new} = x_{id}^{old} + v_{id}^{new} \tag{2}$$

where r_1 and r_2 are random values between 0 and 1. c_1 and c_2 are the learning factors which influence particle movement within a generation. v_{id}^{new} and v_{id}^{old} are respectively the updated velocity and the previous velocity in the d^{th} dimension of the i^{th} particle. x_{id}^{old} and x_{id}^{new} are respectively the current particle position and the updated particle position. w is the inertia weight and it influences the impact of the i^{th} particle in its current velocity and is formulated by Eq. (3).

$$w_{LDW} = (w_{\max} - w_{\min}) \times \frac{Iteration_{\max} - Iteration_i}{Iteration_{\max}} + w_{\min} \tag{3}$$

where $w_{\max}$ is 0.9, $w_{\min}$ is 0.4 and the maximum number of allowed iterations is denoted as $Iteration_{\max}$. The inertia weight can be linearly decreased from 0.9 to 0.4 throughout the iteration process [30]. The global and local searches can be effectively balanced by applying Eq. (3) to the population over the iterations [31].

The pseudo-code of PSO method includes seven steps as shown in Fig. 1. In Step 1, the reasonable values initialize the particles' vectors and define the PSO parameters. In Step 2, the fitness function is used to evaluate the particles' fitness values. In Step 3, the particles' vectors are updated according to each particle's experience (i.e., *pbest*). In Step 4, a best experience (i.e., *gbest*) is updated when the better experience is occurreced. In Step 5, the particle's velocity is updated according to *pbest* and *gbest*. In Step 6, the particle's vector is updated according to current vector and updated velocity. In Step 7, Steps 1 to 6 are repeated until the maximum generation is achieved.