Applied Soft Computing 21 (2014) 415–422

Contents lists available at ScienceDirect

Applied Soft Computing

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

Haplotype inference using a novel binary particle swarm optimization algorithm

Wei Bin*, Zhao Jing

Xi'an Jiatong University, State Key Laboratory for Manufacturing Systems Engineering, School of Electronic and Information Engineering, Xi'an 710049, China

ARTICLE INFO

Article history:
Received 30 August 2011
Received in revised form
14 November 2011
Accepted 22 March 2014
Available online 2 April 2014

Keywords:
Haplotype inference
Pure parsimony
Genotypes
Binary particle swarm optimization

ABSTRACT

The knowledge of haplotypes allows researchers to identify the genetic variation affecting phenotypic such as health, disease and response to drugs. However, getting haplotype data by experimental methods is both time-consuming and expensive. Haplotype inference (HI) from the genotypes is a challenging problem in the genetics domain. There are several models for inferring haplotypes from genotypes, and one of the models is known as haplotype inference by pure parsimony (HIPP) which aims to minimize the number of distinct haplotypes used. The HIPP was proved to be an NP-hard problem. In this paper, a novel binary particle swarm optimization (BPSO) is proposed to solve the HIPP problem. The algorithm was tested on variety of simulated and real data sets, and compared with some current methods. The results showed that the method proposed in this paper can obtain the optimal solutions in most of the cases, i.e., it is a potentially powerful method for HIPP.

1. Introduction

In the post-genome era, as the high-throughput genomic technologies are available, correlating variation in DNA sequence with differences phenotypic (diseases, skin color and so on) has attracted increasing attention [1–3]. Single Nucleotide Polymorphisms (SNPs) are the most common form of DNA variation [4–6]. Studies showed that haplotypes (the combination of SNPs alleles on the same chromosome) can provide more information than genotypes (the conflated data of two haplotypes) in association studies [7–11]. However, technological limitations make it currently impractical to directly collect haplotypes in experimental way [12,13].

There are two ways to solve the Haplotype Inference (HI) problem (find a set of haplotype pairs to explain (or solve) the given genotypes): (1) haplotyping genetically related individuals; (2) haplotyping a population without pedigree information. By the first way, one can get a better estimate of haplotypes, however, it involves significant additional costs [14]. The second one employs computational methods to infer the haplotype from the given genotype data [15–21].

On the basis of the fact that the number of observed distinct haplotypes is vastly smaller than the total number of possible haplotypes [17,22,23], the pure parsimony criteria were proposed [22]. The haplotype inference by pure parsimony (HIPP) is to search the smallest number of distinct haplotypes that can solve the given genotypes [22]. HIPP was proved to be an NP-hard problem [22,24]. Several methods were proposed to solve it [17,25]. The first method was based on Integer Linear Programming (ILP) [26,27]. Recently, Boolean Satisfiability (SAT), Pseudo-Boolean Optimization (PBO), and Answer Set Programming (ASP) were proposed [28–30]. However, most of these methods need to input all of the possible resolutions of each genotype.

As far as we know, the heuristic search algorithm is one of the effective ways to solve the NP-hard problem [31]. Recently, binary particle swarm optimization (BPSO) has been steadily gained attention from the research community because it has many advantages, such as simpler implementation and fewer parameters need to adjust, over other evolutionary algorithms [32]. However, the standard BPSO algorithm usually sinks into the local optimal search space at the later stage of the particles' evolution [33]. In this paper, we propose a heuristic method based on a novel BPSO for the HIPP problem (called NBPSOHAP), and the compatible and incompatible relations between genotypes and haplotypes are used to guide the search direction of NBPSOHAP. We evaluated our algorithm on several simulated and three real data sets, and compared the results with that of some existing methods. The experimental

* Corresponding author. Tel.: +86 13991387365.
E-mail address: weibin82@126.com (B. Wei).

http://dx.doi.org/10.1016/j.asoc.2014.03.034
1568-4946/

results showed that our method can obtain the optimal solutions in most of the cases and outperforms the comparison algorithms.

The rest of the paper is organized as follows. Section 2 introduces the problem of HIPP briefly. Section 3 describes the algorithm proposed in this paper. The dataset and experimental results are provided in Section 4. Finally, we give our conclusions in Section 5.

2. Haplotype inference

One nucleotide of A, C, G or T in the DNA sequence is replaced by any others, e.g., from CCCTAC to CCTTAC, then we call this variation (C→T) as a single nucleotide polymorphism (SNP) [34]. Diploid organisms have two haplotypes, whose positions can be represented by the symbols 0 or 1, where 0 stands for the original base and 1 for the mutant. A position in the genotype has the value of 0 or 1 (called *homozygous*) if the both haplotypes have the same value 0 or 1; otherwise, it is 2 (called *heterozygous*).

Definition 1 *(HI).* Given h genotypes finding a set of haplotypes, such that each genotype is resolved by a pair of haplotypes. A genotype G_i is resolved by a pair of haplotypes H_j, H_k, i.e.:

If $G_{il} = 2$ then $H_{jl} = H_{kl} = 0$

If $G_{il} = 1$ then $H_{jl} = H_{kl} = 1$

If $G_{il} = 2$ then $H_{jl} = 1, H_{kl} = 0$ or $H_{jl} = 0, H_{kl} = 1$

For a genotype with n heterozygous positions, there are $2^n - 1$ possible pairs of haplotypes to resolve it. However, studies showed that the number of observed haplotypes is very small, though genotypes exhibit a great diversity [17,35].

Definition 2 *(HIPP).* Given a set of genotypes, the HIPP aims to find the minimum number of haplotypes to resolve the given genotypes.

3. Methods

3.1. A novel BPSO

In standard BPSO algorithm, each particle is treated equally [36], that is, every particle updates its velocity and position functions according to (1) and (2), respectively [37,38].

$$v_{ij}(t+1) = v_{ij}(t) + c_1 r_{1j}(p_{ij}(t) - x_{ij}(t)) + c_2 r_{2j}(p_{gj}(t) - x_{ij}(t)) \quad (1)$$

$$x_{ij} = \begin{cases} 1 & rand() < S(v_{ij}) \\ 0 & else \end{cases} \quad (2)$$

where x_i is the position of the ith particle; v_i is the velocity of the ith particle; p_i is the best position found by ith particle; p_g is the best position found so far by the entire swarm.

However, this is far from the real situation in a social group. According to sociology, individuals have different status in a social group, and should be differentiated according to the status. Inspired by multi-level organizational learning behavior [39], we propose a novel BPSO (NBPSOHAP) for HIPP problem. In NBPSOHAP, a swarm consists of two types of particles: leaders and followers. We assume that there are K leaders at each iteration and the remaining particles are followers. Due to the fact that leadership roots in the characteristics that certain individuals possess [40], at each iteration, a number of particles, which have better fitness value, are chosen as leaders. The leaders are used to seek the construct of creativity and innovation [41]. In this study, Eqs. (1) and (2) are used to accomplish these tasks.

The status within a group may reflect the importance (weight) of the individual when making a decision [42]. In addition, leaders' decisions are more likely to be accepted by the other members of the group, i.e., followers are likely to make their own decisions mainly based on leaders' [43,44]. Thus, at the tth iteration, the behavior of the followers can be modeled as a random walk toward the leaders.

$$prob_{ij}^F(t+1) = \begin{cases} \min\left(prob_{ij}^F(t) + \frac{1}{\alpha_{LIF}}, 1\right), & if \ \frac{\sum_{j=1}^{K} x_{ij}^L(t)}{K} \geq 0.5 \\ \max\left(0, prob_{ij}^F(t) - \frac{1}{\alpha_{LIF}}\right), & if \ \frac{\sum_{j=1}^{K} x_{ij}^L(t)}{K} < 0.5 \end{cases} \quad (3)$$

$$v_{ij}^F(t+1) = \begin{cases} -V_{\max}, & if \ prob_{ij}^F(t+1) = 0 \\ S^{-1}(prob_{ij}^F(t+1)), & else \\ V_{\max}, & if \ prob_{ij}^F(t+1) = 1 \end{cases} \quad (4)$$

$$x_{ij}^F(t+1) = \begin{cases} 1 & rand() < prob_{ij}^F(t+1) \\ 0 & else \end{cases} \quad (5)$$

where $S^{-1}(x) = \ln(x/(1-x))$; K is the number of leaders; L and F indicate the 'leaders' and 'followers', respectively; and α_{LIF} (named leader impact factor) is an integer which is used to control the converge speed of followers. The smaller α_{LIF}, the faster of 'followers' converge to 'leaders'. Eq. (3) means that if more than half (or exactly half) of leaders choose '1' (at bit j) in tth iteration then the followers' corresponding bit will have a higher probability to be set to '1' in $(t+1)$th iteration.

Lack of the diversity, particularly during the latter stages, is the dominant factor of the convergence of particles to local optimum solutions [45], therefore, the concept of "mutation" is adopted to enhance the global search capability. In this way, when the global optimal solution (p_g) does not improve with the increase of iterations, the following operation is executed:

$$x_{ij}(t+1) = \begin{cases} 1 - x_{ij}(t+1), & if \ rand() < p_m \\ x_{ij}(t+1), & else \end{cases} \quad (6)$$

where p_m is the mutation probability.

According to above analysis, the flow chart of the NBPSOHAP is given in Fig. 1.

3.2. Particle representation

It is trivial to resolve any homozygous genotype vector. However, it is difficult for heterozygous sites.

Example 1. Given a genotype $g = 22101$, the possible haplotype pairs could be represented by the follows:

(1) $h_1 = 11101, h_2 = 00101$
(2) $h_1 = 10101, h_2 = 01101$.

Therefore, only the heterozygous sites need to be dealt with in algorithm. For m genotypes (each one containing n_i heterozygous sites), a particle of NBPSOHAP can be represented by a binary string with $\sum_{i=1}^{m} n_i$ dimensions.

Example 2. Let us consider an HIPP instance constituted by five genotypes: 21101, 21210, 21201, 22120 and 21002:

$$G = \begin{pmatrix} 2 & 1 & 1 & 0 & 1 \\ 2 & 1 & 2 & 1 & 0 \\ 2 & 1 & 2 & 0 & 1 \\ 2 & 2 & 1 & 2 & 0 \\ 2 & 1 & 0 & 0 & 2 \end{pmatrix}$$