Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 38 (2012) 1704 – 1708

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Modeling Optimization and Computing (ICMOC-2012)

An Approach to Frequent Pattern Discovery from Gene Expression Data using PSO Variants

Shruti Mishra[a*], Sandeep Kumar Satapathy[b], Debahuti Mishra[c] and Vinita Debayani Mishra[d]

[a, b, c] *Institute of Technical Education and Research, Siksha O Anusandhan Deemed to be University, Bhubaneswar, Odisha, India*
[d] *NIIS Institute of Business Adminstration, Bhubaneswar, Odisha, India*

Abstract

Pattern mining has always attracted a huge attention for generation of large amount of patterns and association between them. Though it's one of the major data mining tasks but it has always been a time consuming process as a large scale of patterns and associations rules gets generated. To reduce the time of consumption it was preferable to discretize the data matrix in the range of 0 to 1 and for this the fuzzy the membership function has been used which is quite simple in its concept and strategy. Owing to the concept of fuzzy logic, certain evolutionary algorithms (EAs) also gained popularity to optimize the process of mining patterns from the fuzzy sets. For this, Particle swarm optimization (PSO) was used which is supposed to provide better results as compared to other EA like genetic algorithm, ant colony optimization etc. But it was found that there are certain versions of PSO that provided much better results than the standard PSO algorithm. In this paper, the gene expression data set was fuzzified for the purpose of discretization in the range of 0 to 1. A Frequent Pattern (FP) growth algorithm was used to generate set of frequent patterns. These patterns were used as the initial population and the mean squared residue (MSR) score was used as an evaluation criteria. Fully Informed Particle Swarm Optimization (FIPSO), Dynamic Multi Swarm Particle Swarm Optimization (DMS-PSO), Comprehensive Learning Particle Swarm Optimization (CLPSO), Vector Evaluated Particle Swarm Optimization (VEPSO) etc are the certain versions of PSO that were used and they provided much better results as compared to standard PSO algorithm. But the VEPSO algorithm outperformed the other three algorithms in terms of generation of best individual frequent patterns, runtime and the volume of mean squared residue (lower the MSR score the better is the quality of the patterns).

Keywords: Fuzzy logic; Particle Swarm Optimization (PSO); FIPSO; DMS-PSO; CL-PSO; VE-PSO; FP-growth algorithm

1. Introduction

Frequent pattern mining has been a focused theme in data mining research for over a decade. Frequent patterns discovered are basically useful to discover association rules which not only reveal associations between genes and environments to identify gene regulation pathways but also help to uncover gene networks. The discovery of frequent patterns in large scale is always a time consuming work. Fuzzy set theory [1] was basically used to deal with uncertainty, including vagueness and ambiguity. In frequent pattern mining, to reduce the time factor the fuzzy membership function μ = [0, 1] has been used to discretize the matrix in the range of [0, 1].

* Tel.: +91-674-2350181; fax: 91-674-2351880.
E-mail address: shrutimishra@iter.ac.in

1877-7058
doi:10.1016/j.proeng.2012.06.207

PSO [2] is a robust stochastic optimization technique based on the movement and intelligence of swarms. It applies the concept of social interaction to problem solving. In order to improve the performance of PSO, different versions of PSO were introduced called FIPSO [3], VEPSO [4], DMS-PSO [5] and CLPSO [6]. FIPSO was used were the particle uses information from all its neighbours rather than just the best one. It's an alternative that is more concise and promises to perform more effectively than the traditional particle swarm algorithm. The DMS-PSO was constructed based on the local version of PSO with a new neighbourhood topology. PSO with small neighborhoods performs better on complex problems. Hence, to slow down convergence speed and to increase diversity to achieve better results on multimodal problems, in the DMS-PSO, small neighbourhoods are used. The population is divided into small sized swarms. Each sub-swarm uses its own members to search for better regions in the search space. CLSPO [3] is a novel learning strategy that improves the original PSO that is all the particles' *pbest* are used to update the velocity of any one particle. It also ensures that the diversity of the swarm is preserved to discourage premature convergence. One of the most important properties of CLPSO is that it does not introduce any complex operations to the original simple PSO framework. In VEPSO, each swarm is evaluated using only one of the objective functions of the problem under consideration, and the information it possessed for this objective function is communicated to the other swarms through the exchange of their best experience. The best position attained by each particle separately as well as the best among these positions are the main guidance mechanism of the swarms. So, exchanging this information among swarms leads to Pareto optimal points.

In this paper, a gene expression data matrix was considered that was fuzzified in order to range the value in between 0 to 1. Though there are various frequent pattern mining algorithms available but the FP growth algorithm was used in order to generate a set of frequent patterns which was considered as an initial population for the algorithm. It was observed that the VEPSO algorithm generated some of the best individual frequent patterns than the FIPSO, DMS-PSO and CLPSO.

The layout of the paper is as follows: section 2 deals with related work based on FIPSO, DMS-PSO, CLPSO and VEPSO. Section 3 gives the work plan model, section 4 states the experimental evaluation and finally section 5 deals with the conclusion and future work.

2. Related Work

Montes *et al*. [7] studied the convergence behavior of the particles when using topologies with different levels of connectivity. They also showed that the particles tend to search a region whose size decreases as the connectivity of the population topology increases. Liang *et al.* [5] proposed a DMS-PSO with local search for solving the CEC 2008 large scale global optimization problem. Here the population of the DMS-PSO was divided into many small sub swarms and these sub swarms were regrouped frequently to exchange the information among all the particles. By combining the exploration and the exploitation together the neighborhood structure gave better performance on complex problems. Tang *et al.* [6] presented a CLPSO strategy for structural parameter estimation which ensured that the diversity of the swarm is preserved to discourage premature convergence. It has been observed that the CLPSO outperforms the PSO algorithm on no prior knowledge case and significantly improves the results on partial output search. Vlachogiannis *et al.* [4] used a parallel VEPSO approach and applied it to reactive power control of power systems in steady state. The results showed that the approach was efficient for solving the multi-objective problem of reactive power control and when compared with other evolutionary techniques it outperformed in the computing time.

3. Proposed Model

Fig.1. shows our proposed model were the Leukaemia [13] data set has been considered which was fuzzified by using fuzzy membership function that was further categorized into low and high set. A FP

growth algorithm was used which is one of the most popular frequent pattern mining algorithm available; the MSR score was calculated which acted as a selection criteria for the purpose of evaluation and then the different versions of PSO algorithm was used to generate the best individual patterns.

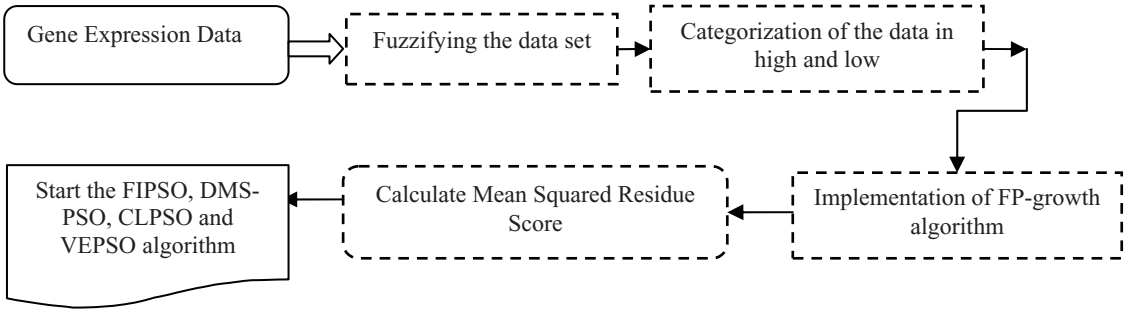

Fig.1. Schematic representation of proposed model

4. Experimental Evaluation

Step-I: Fuzzifying the Original Matrix

Leukemia [13] data set of 72 * 50 was considered and was fuzzified using the triangular membership function to substitute the element values in the range of [0, 1] in order to discretize the data set.

Step-II: Division of Fuzzy data matrix, calculation of mean squared residue

The fuzzy data matrix is categorized into low and high sets which depict the under-expressed and over-expressed genes. Also the mean squared residue value was calculated which acts as a selection criteria. Though other parameters can be used as the selection criteria but for our domain it gave us a better result. The set of frequent patterns generated from the FP-growth algorithm is considered as the set of initial population. The evaluation criteria for the above algorithm can be stated as follows:

$$fitness(fp) = \frac{fp * mean\ squared\ residue\ score}{number\ of\ fp} \quad (1)$$

In the following steps, the variants of PSO have been implemented by considering the initial population and fitness (*fp*) (as shown in (1)) with the same value as discussed above.

a) *FIPSO*: In FIPSO, particles or patterns uses the information provided by all its neighbors in order to update its velocity. The velocity and position update rule is:

$$v_{i,j}^{t+1} = \chi \left[v_{i,j}^{t} + \frac{1}{K_i} \sum_{n=1}^{K_i} \mathbb{U}(0, \varphi) \left(p_{N_{i(n),j}}^{t} - x_{i,j}^{t} \right) \right] \quad (2)$$

Where, χ = constriction factor; K_i =No. of particles or patterns in the neighborhood of particle *i*; $\mathbb{U}(0, \varphi)$= uniformly distributed random number in the range of $(0, \varphi)$ where φ is an acceleration coefficient; N_i (n) = function that returns the index of the n^{th} neighbor of pattern *i* and $p_{N_{i(n),j}}^{t}$= j^{th} component of the previous best position of the n^{th} neighbor of *i*. Here, the parameter values were taken as: $= 0.7298$;$\varphi = 4.1$; Population size = 40.

b) *DMS-PSO*: This algorithm [8] has two phases: one in which the frequent patterns or the populations are divided into the small sized swarms and the other being the randomly regrouping schedule where maximum information exchange among the patterns are allowed in order to enhance the diversity of the pattern. Parameter values taken are number of swarms (n) = 30, each swarm population (ns) = 3 and the regrouping period (R) =100.

c) *CLSPO*: In this algorithm [9], the velocity updation of the particles was basically done using the formula given in (3):

$$v_i^d = w * v_i^d + c * rand_i^d * \left(pbest_{fi(d)}^d - X_i^d\right) \quad (3)$$

Where, v is the velocity; f_i defines which pattern's *pbest* should the pattern i follow, $pbest_{fi(d)}^d$ is the dimension of any pattern's *pbest* including its own *pbest* denoted as P.

d) *VEPSO*: Here, the algorithm [10] initializes randomly the two swarms (M_1, M_2) within the feasible solution space and the time is set to, t=0. Each particle in the initial population is evaluated until it satisfies the constraints and the ring migration topology is used for the global best position and the best previous position of the pattern (particles) in the s^{th} swarm for the evaluation of the velocities of the j^{th} swarm. The parametric values: number of swarms, M=2; number of particles in each swarm = 12; maximum number of allowed iterations =200; cognitive parameters, c_1=0.4; social parameters, c_2=0.4; inertia weight, W_{min} = 0.2 and W_{max}=1.

The result of the above four algorithms in comparison to the standard PSO [11] is shown in table 1 and fig. 2. It was observed that the VEPSO algorithm was much better than the standard PSO algorithm for fuzzy frequent patterns.

Table 1. Pattern comparison between Standard PSO and its versions

Algorithm	Mean Squared Residue	Patterns generated	Average Runtime (in milliseconds)
Standard PSO-fuzzy FP growth	139.90	520	2865
FIPSO- fuzzy FP growth	138.06	532	2797
DMS-PSO fuzzy FP growth	138.05	549	2687
CLPSO fuzzy FP growth	137.89	554	2678
VEPSO fuzzy FP growth	137.01	562	2590

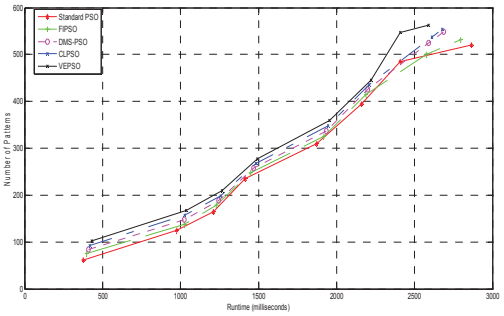

Fig. 2. Result of comparison for the above two algorithms

5. Conclusion

The FIPSO algorithm is very sensitive to changes in the population topology. The velocity update rule plays a major criterion in FIPSO that considers all the neighbors of a particle to update its velocity instead of just the best one neighbor as in standard PSO algorithm. But the VEPSO algorithm was found to be better as compared to the DMS-PSO, CLPSO and FIPSO algorithms because here the generations of best individual frequent patterns were more and of better quality (assessed by the MSR score). Also the average runtime of the algorithm was found to be good.

References

[1] Zadeh HJ. Fuzzy Set Theory and its application. *Kluwer Academic Publisher*; 1991.
[2] Kennedy J and Eberhart RC. Particle Swarm Optimization. *Proc. of Int. Conf. on Neural Networks*; 1995; 4: 1942-1948.
[3] Mendes R, Kennedy J and Neves J. The fully informed particle swarm: Simpler, maybe better. *IEEE Transactions on Evolutionary Computation*; 2004; 8(3): 204-210.
[4] Vlachogiannis JG, Lee KY. Multi-objective based on parallel vector evaluated particle swarm optimization for optimal Steady-state performance of power systems. *Expert Systems with Applications*; 1995; 36: 10802-10808.
[5] Zhao SZ, Liang JJ, Suganthan PN, Tasgetiren MF. Dynamic Multi-Swarm Particle Swarm Optimizer with Local Search for Large Scale Global Optimization. *IEEE World Congress on Computational Intelligence*; 2008: 3845-3852
[6] Tang H, Zhang W, Fan C, Xue S. Parameter Estimation Using a CLPSO strategy. *IEEE Congress on Evolutionary Computation* (CEC); 2008: 70-74.
[7] Marco A, Montes de Oca and Stützle T. Convergence behavior of the fully informed particle swarm optimization algorithm. *Proc. of Int. Conf. on genetic and evolutionary computation*; 2008: 1-16.
[8] Mishra S, Mishra D, Satapathy SK. Fuzzy Frequent Pattern Mining from Gene Expression Data using Dynamic Multi-Swarm Particle Swarm Optimization. *2nd Int. Conf. on Computer, Communication, Control and Information Technology*; 2012 (To be published in *Procedia Technology*, ISSN: 2212-0173 2012) (Accepted)
[9] Mishra S, Mishra D, Satapathy SK. CLPSO- Fuzzy Frequent Pattern Mining from Gene Expression Data. *2nd Int. Conf. on Computer, Communication, Control and Information Technology*; 2012 (To be published in *Procedia Technology*, ISSN: 2212-0173 2012) (Accepted)
[10] Mishra S, Mishra D, Satapathy SK. Fuzzy Frequent Pattern Mining from Gene Expression Data using Vector Evaluated Particle Swarm Optimization. *Int. Conf. on Information and Education Technology*; 2012 (To be published in Procedia Technology, ISSN: 2212-0173) (Accepted)
[11] Mishra S, Mishra D, Satapathy SK. Particle Swarm Optimization based Fuzzy Frequent Pattern Mining from Gene Expression Data. *Int. Conf. on Computer and Communication Technology*; 2011:15-20.
[12] *www.ucirepository.com*