CLPSO - Fuzzy Frequent Pattern Mining from Gene Expression Data

Shruti Mishra⁠¹, Sandeep Ku. Satapathy⁠², Debahuti Mishra⁠³

Institute of Technical Education and Research, Siksha O Anusandhan Deemed to be University, Bhubaneswar, Odisha, India

Abstract

Frequent pattern mining has attracted much attention and wide applications owing to its simple concept and strategy. It is one of the most important tasks in data mining and knowledge discovery. But usually a large number of frequent patterns get generated from a large scale of data matrix which is a time consuming affair. So, in order to discretize the data matrix a mathematical concept called fuzzy logic was used. It generalizes the data matrix values in the range of 0 to 1. In the due course of time, an evolutionary algorithm, called Particle swarm optimization (PSO) has also gained much popularity. But due to the premature convergence of PSO, a comprehensive learning strategy was introduced that used all particles’ best information to update a particle’s velocity. It also enabled the diversity of the swarm to be preserved to discourage premature convergence. In this paper, frequent patterns were generated from the fuzzy dataset (data matrix converted into fuzzy data matrix) using the Frequent Pattern (FP) growth algorithm. In order to generate some of the best individual frequent patterns out of the entire set of patterns, the CLPSO algorithm was used with a selection measure called mean squared residue (MSR) score. It was noted that the CLPSO algorithm outperformed the traditional PSO algorithm in the generation of the best individual patterns with a comparatively lower MSR value.

Keywords: Frequent pattern mining; Fuzzy logic; Particle Swarm Optimization; Comprehensive Learning Particle Swarm Optimization

1. Introduction

Data mining has recently attracted considerable attention from database practitioners and researchers as it has been applied to many fields such as decision support, medical diagnosis etc. Within that, frequent pattern mining plays an important role in data mining research areas especially in the real time data mining research and have been successfully applied in business and scientific data for discovering interesting association patterns. As such, it has become a promising strategy in analyzing the gene expression set. However, discovering frequent patterns in large scale data is an extremely time consuming task. Fuzzy set theory that was proposed by Zadeh [1] is primarily concerned with quantifying and discretizing the data into an acceptable range of 0 and 1.
As an evolutionary computation technique, PSO has attracted much attention and wide applications, owing to its simple concept, easy implementation and quick convergence [2]. In order to improve the PSO’s performance, an improved PSO was used by incorporating a comprehensive learning strategy into the original PSO to enhance convergence properties. It was known as CLPSO. CLSPO [3] is a novel learning strategy that improves the original PSO that is all the particles’ \textit{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.

Though there are various frequent pattern mining techniques available, in this paper the FP growth algorithm was used on a fuzzy data matrix to generate the frequent patterns. In order to generate some of the best frequent patterns, an evolutionary algorithm called CLPSO algorithm was used that uses the mean squared residue score (MSR) as the selection criteria. The lower the MSR value the better is the quality of the patterns obtained.

The layout of the paper is as follows: section 2 deals with related work based on CLPSO. Section 3 gives the work plan model, section 4 describes the experimental evaluation and result analysis and finally section 5 gives the conclusion and future directions.

2. Related Work

Tang et al. [4] presented a CLPSO strategy for structural parameter estimation which ensured that the diversity of the swarm is preserved to discourage premature convergence. It was also observed that the CLPSO outperforms the PSO algorithm on no prior knowledge case and significantly improves the results on partial output search. Hamidi et al. [3] used the CLPSO technique for finding the optimal fuzzy rules and membership functions for segmentation of color images. Wu et al. [5] proposed a novel scheme that they called as the opposition based comprehensive learning particle swarm optimizers that employed opposition based learning for population initialization and also for exemplar selecting. Liang et al. [6] proposed novel constraint-handling mechanism employed in CLPSO to handle constrained real-parameter optimization problems. With the help of the new constraint-handling mechanism, the particles were adaptively assigned to explore different constraints in the search process. In order to improve the search efficiency and global search ability, a co-evolutionary schedule was also used.

3. Proposed Model

Fig 1 shows the proposed model where the gene expression data matrix is fuzzified using the fuzzy framework. Then various frequent pattern mining algorithms are used to generate frequent patterns and then using the mean squared residue score as the selection measure we begin the CLPSO algorithm.
4. Experimental Analysis and Evaluation

To achieve the best individual frequent patterns, the experimental evaluation has been categorized into five basic stages:

**Stage I: Implementation of fuzzy framework on the original dataset:** The original gene matrix dataset was considered where the triangular membership method was used to fuzzify the dataset in order to discretize it in the range of 0 to 1 (as shown in fig 2).

![Fig. 2. Example of a data matrix and the fuzzified matrix set of 3 genes and 3 conditions](image)

**Stage II: Categorization of the fuzzified dataset:** The fuzzify gene matrix dataset was divided and categorized into two sets called as low and high where the low set indicate the under-expressed genes and the high set consider the over-expressed genes.

**Stage III: Use of frequent pattern mining algorithms:** Various frequent pattern mining algorithms like apriori, vertical data format and FP-growth were used to generate frequent patterns [7]. These patterns were used as the initial set of population in the CLPSO algorithm.

**Stage IV: Calculation of Mean Squared Residue (MSR):** Here, the mean squared residue (MSR) was calculated and was used as a selection measure. The fitness function evaluation was basically done using this MSR value.

**Stage V: Implementation of Comprehensive Learning Particle Swarm Optimizer (CLPSO):**

(a) **Parameters Undertaken**
- No. of swarms \( (n) = 30 \)
- Population Size = 30
- Dimension, \( D = 1 \)
- Refreshing gap = 0.8

The calculation for \( P_c \) (the learning probability) was done using the generalized formula (1):

\[
P_c = 0.05 + 0.45 \times \left( \frac{\exp \left( \frac{\text{fitness}(fp)}{\text{MSR}} \right)}{\exp \left( \frac{\text{MSR}}{\text{fitness}(fp)} \right)} \right)
\]  

(1)

(b) In this algorithm the velocity updation of the particles was basically done using the formula given in (2):

\[
v^p_i = \Delta v_i + \epsilon \times \tau \times \alpha \times (p_{bestf_i(d)} - X^p_i)
\]

Where, \( v \) is the velocity, \( f_i \) defines which pattern’s \( p_{best} \) should the pattern \( i \) follow, \( p_{bestf_i(d)} \) is the dimension of any pattern’s \( p_{best} \) including its own \( p_{best} \), \( P \).

(c) The fitness function is evaluated as follows in (3):

\[
\text{fitness}(fp) = \frac{\text{mean squared residue score}}{\text{number of } fp}
\]

where, \( fp \) are the frequent patterns.

(d) For each pattern \( i \) a random number is always generated.

(e) If random number > \( P_c \),

then

Learn from own \( p_{best} \).
else
   Learn from another pattern’s pbest

(f) When the pattern’s dimension learns from another pattern’s pbest then the tournament selection procedure is followed:
   - Randomly choose two patterns from the population which excludes the pattern whose velocity has to be updated.
   - The fitness values of the two patterns pbests were compared and the better one was selected.

Then the winner’s pbest was used to set an exemplar to learn from that dimension.

The results obtained after implementing the above CLPSO based Fuzzy FP growth algorithm yields us much better desired result as compared to the PSO based fuzzy FP growth algorithm [7] as shown in table 1 and fig 3. The accuracy of finding the best patterns was much more in CLPSO than in traditional PSO.

Table 1. Comparison based on the parameters mentioned

| Algorithm              | Mean Squared Residue | Patterns generated | Runtime (in milliseconds) | Accuracy of finding best individual patterns |
|------------------------|----------------------|--------------------|---------------------------|---------------------------------------------|
| CLPSO based Fuzzy FP growth | 137.89              | 554                | 2678                      | 90.61%                                      |
| PSO based Fuzzy FP growth     | 139.90              | 520                | 2865                      | 88.01%                                      |

Fig. 3. Graphical results of the comparison of the two algorithms

Comparison for the apriori algorithms, vertical data format, FP growth, PSO based FP growth using the fuzzified dataset including the CLPSO based fuzzy FP growth is shown in table 2 and fig 4. It has been found that out of all the above mentioned algorithms the CLPSO based fuzzy FP growth performs better.

Table 2. Comparison of the five algorithms based on the parameters

| Algorithm              | Average Mean Squared Residue | No. of Frequent patterns generated | Runtime (milliseconds) |
|------------------------|------------------------------|-----------------------------------|------------------------|
| Fuzzy Apriori algorithm | 205.01                       | 323                               | 3455                   |
| Fuzzy Vertical data format | 174.20                       | 365                               | 3233                   |
| Fuzzy FP growth        | 159.80                       | 415                               | 3010                   |
| PSO based Fuzzy FP growth     | 139.90                       | 520                               | 3455                   |
| CLPSO based Fuzzy FP growth     | 137.89                       | 554                               | 2678                   |
5. Conclusion

CLPSO being a novel learning strategy improved the original PSO were all the particles (patterns here) pbest were used to update the velocity of any one particle. In this paper, we used the CLPSO strategy on the frequent patterns to generate some of the best patterns. From the results, has been observed that CLPSO outperforms the traditional PSO algorithm in terms of generation of the number of frequent patterns, runtime and accuracy of generating the best individual patterns.

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. Hamidi M, Borji A. Color Image Segmentation with CLPSO based Fuzzy. International Journal of Computer Science and Network Security (IJCSNS); 2007; 7(6): 215-221.
4. Tang H, Zhang W, Fan C, Xue S. Parameter Estimation Using a CLPSO strategy. IEEE Congress on Evolutionary Computation (CEC); 2008: 70-74.
5. Wu Z, Ni Z, Zhang C, Gu C. Opposition based Comprehensive Learning Particle Swarm Optimization. Proc. of Int. Conf. on Intelligent System and Knowledge Engineering; 2008: 1013-1019.
6. Liang JJ, Zhigang S, Zhihui L. Co-evolutionary Comprehensive Learning Particle Swarm Optimizer; IEEE Congress on Evolutionary Computation (CEC); 2010: 1-8.
7. 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.