

Chapter 2

The Combination of Evolutionary Algorithm Method for Numerical Association Rule Mining Optimization

Imam Tahyudin and Hidetaka Nambo

Abstract The numerical problem of association rule mining is an updated issue. Numerous authors propose some methods to solved it. A number of them are using the optimization approach by Particle Swarm Optimization (PSO). The problem is that the PSO trapped in local optima when searched the best particle in every iteration. Many researchers solved this problem by combining with Cauchy distribution because it is tremendous for searching in a large neighborhood. Hence, that combination will be implemented to accomplish the numerical association rule mining problem for some objective functions such as confidence, comprehensibility, interestingness. Based on the result the multi-objective of PSO for Numerical Association Rule Mining Problem with Cauchy Distribution (PARCD) showed the better result than the method of Multi-objective Particle Swarm Optimization for Association Rule Mining (MOPAR).

Keywords PSO · Cauchy distribution · Numerical association rule mining · Multi-objective functions

2.1 Introduction

Association rule mining is one of methods in data mining which interesting to discuss deeply. This method mines the data that emerge frequently in the same time together. This method evolves increasingly and it is combined by other multidiscipline like machine learning and evolutionary algorithm. For instance, combination ARM with fuzzy concept [1], ARM with PSO [2] and ARM with GA [3].

I. Tahyudin (✉) · H. Nambo

Graduate School of Natural Science and Technology Division of Electrical Engineering
and Computer Science, Kanazawa University, Kanazawa, Japan
e-mail: imam@blitz.ec.t.kanazawa-u.ac.jp

© Springer Science+Business Media Singapore 2017
J. Xu et al. (eds.), *Proceedings of the Tenth International Conference
on Management Science and Engineering Management*, Advances in Intelligent
Systems and Computing 502, DOI 10.1007/978-981-10-1837-4_2

The familiar algorithms which are used in this methods are A priori and FP growth algorithm. Both of them implemented in specific case like A priori algorithm which used for large number database but the FP growth preferred used for small number database [4]. In addition, both of them appropriate used for categorical data type like gender or binary form while if the data is numerical type such as age, weight or length, it there is additional discretization step which transform the data into categorical type. In contrast, this step has many weaknesses like missing many information and need more time to process it [2, 5].

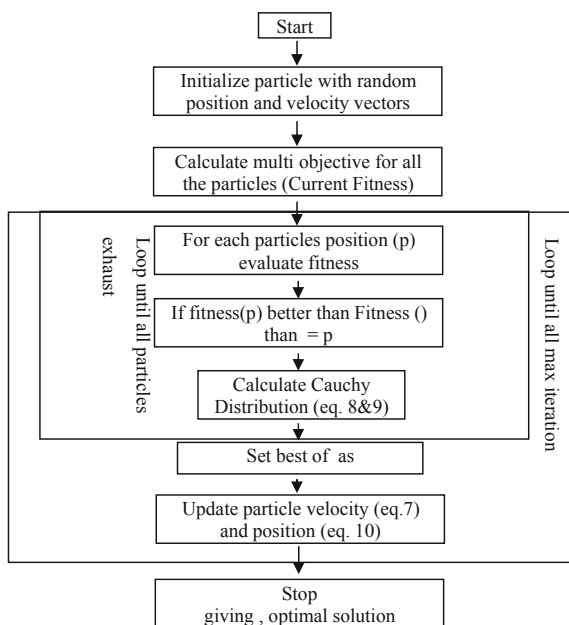
To solve numerical data type in ARM has tried by some researchers. They tried by mono objective just using two parameters which are support and confident. On the other hand, they using multi objectives measurements, not only both parameters but also comprehensibility and interestingness. Moreover, some of them have used pareto optimality for fitness computation while did not use it [4]. One of contribution of this research is combining one of optimization method in numerical ARM, PSO, with Cauchy distribution which was introduced by [6] as the method simple and robust. This combination to prevent the weakness of PSO that premature in searching optimal solution. Hence, this research has aim to bring the hybrid of PSO with Cauchy distribution to solve multi objective numerical ARM optimization by pareto optimality.

The paper is organized as follows: Sect. 2.2 reviews the literature of recent research works; Sect. 2.3 presents the proposed method of combination of PSO and Cauchy distribution for numerical association rule mining optimization (PARCD); Sect. 2.4 gives a discussion and analysis of numerical experiments results for some multi objective problem such as support, confidence, comprehensibility, interestingness, amplitude and coverage; finally, the conclusion and future work are given in Sect. 2.5.

2.2 Literature Review

Nowadays, numerical ARM problem updated to discuss. Many researchers solved this problem in numerous approaches for example by evolutionary algorithm like particle swarm optimization and other machine learning methods like fuzzy and genetic algorithms. In [5] depicts the numerical association rule problem be able to solved by discretization, distribution and optimization. The discretization is done by partitioning and combining, clustering and fuzzy [1, 7]. Then, the optimization is approached by optimized Association rule mining [3], differential evolution [8], Genetic Algorithm (GA) [5, 9, 10] and Particle swarm optimization (PSO) [2, 11, 12]. Whole of them are described in Fig. 2.1.

By those solution, the numerical data can be solved to attain the important information without discretization process [2, 13] and in some method can automatically determine the minimum support and minimum confident based on the optimal threshold without decide by the authors [3, 10].

Fig. 2.1 Flowchart of PARCD

According to the latest paper by [2] that the numerical ARM optimization problem has solved by using PSO well. The strengths of PSO are it can define the parameter without specifying upfront the minimum support and confident and also it able to generate best rule independent of length of frequent item set [14]. On the other hand, PSO method has the weakness like the user has to specify the number of best rule and the time of complexity [14] and also it is not robust in large data [6]. So that, it is still potential to be evolved. One of the ways to diminish the weakness is revealed in [6], that the combination of PSO with Cauchy has approved can rise the leverage of result because the mutation process can reach wider and appropriate to a large database. In other research [15] that combination of them have ability to optimize two-stage reentrant flexible flow shop with blocking constrain. In addition this combination can improve the make span solution by average 15, 60 % and then the performance of this combination higher than HGA [15]. Then, this combination have used to optimize the integration of process planning and scheduling (IPPS) and the result shows the effectiveness of the proposed IPPS method and the reactive scheduling method [16]. This hybrid method has developed by Gen et al. to increase the wide search space in mutation process by using Cauchy distribution like revealed in [15, 17], the result shows that the method can enhancing the evolutionary process with the wide search space. Hence, refer to those previous researches, we have novelty to bring this modification method in [15, 17] to be implemented in numerical association rule mining optimization.

2.3 Proposed Method

2.3.1 Objective Design

In this study Authors use some objective parameters which are support, confidence, comprehensibility, and interestingness. The support criterion measures the ratio of transactions in D containing X , or $\text{sup}(X) = |X(t)|/|D|$. The support of the rule $X \rightarrow Y$ is computed using the following equation

$$\text{Support}(X \cup Y) = |X \cup Y|/|D|. \quad (2.1)$$

This support measure is used for determining the confidence criterion. The confidence measures the quality of rule based on the number of transaction of an AR in the whole dataset. The rule which often emerge in every transaction is considered to have a better quality [2].

$$\text{Confidence} = \text{Support}(X \cup Y)/\text{Support}(X). \quad (2.2)$$

Individually, this confidence measure not be guarantee obtaining the appropriate AR. In order to gain the appropriate coverage and reliability, the resulted rule also be comprehensible and interesting. According to a research, the less number of conditions in the antecedent part of a rule would be, the more comprehensible or understandable is that rule [18]. Hence, the comprehensibility can be measured as below

$$\text{Comprehensibility} = \log(1 + |Y|)/\log(1 + |X \cup Y|), \quad (2.3)$$

where $|Y|$ is the number of rule in the consequent the and $|X \cup Y|$ is the total rule in the consequent and antecedent rules.

The interestingness criterion is used for obtaining hidden information by extracting of such surprising rules. This criterion based on support count of both antecedent and consequent part [18]. The equation is shown in Eq. (2.6).

$$\text{Interestingness} = \left[\frac{\text{Support}(X \cup Y)}{\text{Support}(X)} \right] \times \left[\frac{\text{Support}(X \cup Y)}{\text{Support}(Y)} \right] \times \left[1 - \frac{\text{Support}(X \cup Y)}{|D|} \right]. \quad (2.4)$$

The interestingness formula consists of three parts. Firstly, $[\text{Support}(X \cup Y)/\text{Support}(X)]$, the generation probability of the rule is computed in terms of the antecedent part of the rule. Secondly, $[\text{Support}(X \cup Y)/\text{Support}(Y)]$, the generation probability of the rule is computed in terms of the consequence part of the rule. Finally, in the third part, $[1 - \text{Support}(X \cup Y)/|D|]$, the section $\text{Support}(X \cup Y)/|D|$ shows the probability of generating the rule according to the total records of the dataset ($|D|$). So, its complement, $[1 - \text{Support}(X \cup Y)/|D|]$, means the proba-

bility of not generating the rule. Therefore, a rule with a high value of support count will be considered as a less interesting rule [2].

2.3.2 PSO

Kennedy and Eberhart [12] explained the Swarm Intelligence (SI) that it is an innovative distributed intelligent paradigm for solving optimization problems that originally took its inspiration from the biological examples by swarming, flocking and herding phenomena in vertebrates. In 1995, They found the PSO which incorporates swarming behaviors observed in flocks of birds, schools of fish, or swarms of bees, and even human social behavior.

The main concept of PSO is initialized with a group of random particles (solutions) and then searches for optima by updating generations. During all iterations, each particle is updated by following the two “best” values. The first one is the best solution (fitness) it has achieved so far. This value is called “pBest”. The other “best” value that is tracked by the particle swarm optimizer is the best value obtained so far by any particle in the population. This best value is a global best and is called “gBest”. After finding the two best values; each particle updates its corresponding velocity and position [12].

Each particle p , at some iteration t , has a position $x(t)$, and a displacement velocity $v(t)$. The personal best (pBest) and global best (gBest) positions are stored in the associated memory. The velocity and position are updated using Eqs. (2.5) and (2.6) respectively [6, 12].

$$v_i^{\text{new}} = \omega v_i^{\text{old}} + c_1 \text{rand}().(\text{pBest} - x_i) + c_2 \text{rand}().(\text{gBest} - x_i), \quad (2.5)$$

$$x_i^{\text{new}} = x_i^{\text{old}} + v_i^{\text{new}}, \quad (2.6)$$

where ω is the inertia weight; v_i^{old} is particle velocity of the i th particle before updating; v_i^{new} is particle velocity of the i th particle after updating; x_i is the i th, or current particle; i is the particle's number; $\text{rand}()$ is a random number in the range (0, 1); c_1 is the individual factor; c_2 is the societal factor; pBest is the particle best; gBest is the global best. Particles velocities on each dimension are clamped to a maximum velocity V_{max} [6, 12].

2.3.3 PSO for Numerical Association Rule Mining with Cauchy Dist (PARCD)

Cauchy distribution is used for solving the problem of PSO because of not yield good solution for large scale problems including high dimensional variables. Therefore, make new mutation operation by using the effective particles moving. Kaji in [6, 17]

proposed a Cauchy particle swarm optimization for solving multi-modal optimization problem.

PARCD is the proposed method to implement for solving numerical association rule mining problem. This combination generates the best result because the weakness of PSO has solved by using Cauchy distribution so can prevent the trap of local optima in best particle to gain the global best by long jump searching using Cauchy mutation.

$$v_i(t+1) = \omega(t)v_i(t) + c_1 \text{rand}(\cdot)(p\text{Best} - x_i(t)) + c_2 \text{rand}(\cdot)(g\text{Best} - x_i(t)), \quad (2.7)$$

$$u_i(t+1) = \frac{(v_i(t+1))}{\sqrt{v_{i1}(t+1)^2 + v_{i2}(t+1)^2 + \dots + v_{iK}(t+1)^2}}, \quad (2.8)$$

$$s_i(t+1) = u_i(t+1) \times \tan\left(\frac{\pi}{2} \times \text{rand}[0, 1)\right), \quad (2.9)$$

$$x_i(t+1) = x_i(t) + s_i(t+1). \quad (2.10)$$

2.3.4 Pseudocode of PARCD

Pseudocode of PSO [17]:

Procedure: Combination of PSO and Cauchy distribution for Numerical

ARM Input: PARCD parameters

Output: Multi-objective results

```

Begin
  t ← 0
  initialize  $x_i(t)$  by encoding routine; calculate the multiple objectives for all particles // current fitness
  evaluate  $x_i(t)$  by decoding routine and keep the best solution;
  while (not terminating condition) do
    for each particle  $x_i$  in swarm do
      update velocity  $v_i(t+1)$  // using (2.7)
      update position  $x_i(t+1)$  //using (2.10)
      calculate  $u_i(t+1)$  and  $s_i(t+1)$  //By Cauchy distribution using (2.8) and (2.9)
      evaluate  $x_i(t+1)$  //using (2.10)
      if  $f(x_i(t+1)) < f(p\text{Best}(t))$  then //pbest(t) : historical best position
        update pBest(t) =  $x_i(t+1)$ ; //update the best local position
      end;
      gbest(t+1) = arg min{ $f(p\text{best}(t), g\text{best}(t))$ } //update the global best position
      t ← t + 1;
    end;
  output : the best solution gBest;
end;
```

Table 2.1 Properties of the datasets

Dataset	No. of records	No. of attributes
Quake	2178	4
Basketball	96	5
Body fat	252	15
Pollution	60	16
Bolt	40	8

Table 2.2 The parameter setup

Parameter	Population size	External repository size	Number of iteration	C_1 and C_2	W	limit	xRank
Average	40	100	2000	2	0.63	3.83	13.33

2.4 Experiments and Discussion

2.4.1 Experimental Setup

This research uses benchmark datasets from Bilkent University Function Approximation Repository. There are five data set which are used, Quake, Basketball, Body fat, pollution and Bolt (Table 2.1). This experiment is conducted on 2.7 GHz Intel Core i5, 8 GB main memory, running by Windows 7 and to process the algorithms by using Matlab software.

Firstly, setting up the parameter of some values in the proposed algorithm such as parameter, population size, external repository size, number of iteration, the value of c_1 and c_2 , ω , velocity limit and xRank. They are average, 40, 100, 2000, 2, 0.63, 3.83, and 13.33 respectively. This parameter referred to the previous research by [2] (Table 2.2).

2.4.2 Experiments

Basically, the association rule analysis contains two steps, firstly to determine the frequent item set including the antecedent or consequent from each attributes. Secondly, to implement the proposed algorithm. In this research uses the development method of multi-objective particle swarm optimization of numerical association Rule (MOPAR) which is combined with Cauchy distribution (mutation). We call as PARCD (Particle swarm optimization of numerical association rule with Cauchy distribution).

The Table 2.3 shows about the comparison of support value between PARCD and MOPAR method. Generally, the support percentage of PARCD method is better

Table 2.3 The comparison of support value

Dataset	Support (%)	
	PARCD	MOPAR
Quake	22.97	46.26
Basket ball	61.04	32.13
Body fat	73.94	10.1
Bolt	250.84	107.29
Pollution	60.45	52.14

Table 2.4 The Comparison of number of values and confidence values

Dataset	Number of rules		Confidence (%)	
	PARCD	MOPAR	PARCD	MOPAR
Quake	51	57	86.73 \pm 25.88	82.31 \pm 28.91
Basket ball	78	84	92.69 \pm 17.87	92.67 \pm 16.65
Body fat	32	29	81.26 \pm 30.67	43.59 \pm 61.15
Bolt	42	39	96.88 \pm 9.49	88.91 \pm 9.49
Pollution	12	2	34.96 \pm 43.91	23.02 \pm 40.04

than MOPAR method. Only one dataset that the value is opposite unlike the other datasets; Quake, which is the value in PARCD method is almost half value from MOPAR methods. The considerably highest gap percentage support value is Bolt dataset which is just over by 150 %. On the other hand, the lowest one is Body fat dataset which is approximately one per seven from the percentage its result by PARCD. In addition, the percentage of Basket ball by the proposed method doubled from the value by MOPAR. Interestingly, the value of pollution dataset is nearly similar just about of 7 %.

This result makes the argue that the combination of PSO with Cauchy distribution can reach the large space to search the optimal value. This result also affects to the value of rule numbers and the confidence percentage which are generated in Table 2.4.

This table reveals the differentiation both of PARCD and MOPAR method of number of rules and confidence value. According to number of rules, there are three datasets which are the PARCD method is better than MOPAR method. They are body fat, bolt and pollution. However, it clearly shows in confidence value that all the value in PARCD method is higher than the MOPAR method. Although, the highest value of rules number, basket ball, is higher at 6 from the opponent of MOPAR, the confidence value is almost the same even little bit higher. Interestingly, the lowest value of rule number also by MOPAR method which is pollution (about one sixths from the PARCD method). Furthermore, the highest percentage of confidence is Bolt which is at 96,88 %. In addition, the percentage of body fat dataset by PARCD method is two times higher than its value in MOPAR method and then for quake and pollution datasets are higher by about 4 and 10 % respectively.

Table 2.5 The comparison of comprehensibility value

Dataset	Comprehensibility (%)	
	PARCD	MOPAR
Quake	785.2 \pm 37.72	786.14 \pm 419.67
Basket ball	545.80 \pm 167.74	424.65 \pm 192.63
Body fat	333.49 \pm 218.95	204.87 \pm 235.46
Bolt	231.08 \pm 168.35	271.25 \pm 168.35
Pollution	110.63 \pm 165.76	65.82 \pm 130.49

Table 2.6 The comparison of interestingness value

Dataset	Interestingness (%)	
	PARCD	MOPAR
Quake	2.34 \pm 9.30	4.67 \pm 11.40
Basket ball	6.56 \pm 21.16	4.99 \pm 5.18
Body fat	10.61 \pm 21.03	21.71 \pm 9.30
Bolt	43.43 \pm 39.68	23.70 \pm 39.68
Pollution	9.51 \pm 18.61	10.23 \pm 27.88

The confidence value contains the average of iteration and the standard deviation is to show that the final value is acceptable from fluctuation in some iteration and the stable behavior in every run [2]. It also done in comprehensibility value (Table 2.5), interestingness value (Table 2.6).

This Table 2.5 obviously shows that three kinds of dataset by PARCD method have the higher value than by MOPAR methods. They are basket ball, body fat and pollution. The highest value of two methods is nearly same, it is Quake dataset (about 785). In contrast the lowest one is pollution dataset which is at 110.63 and 65.82 % from PARCD and MOPAR method respectively. The basket ball and body fat dataset by PARCD method are higher about 100 % than their value in MOPAR method. Moreover, the bolt dataset is little bit less than approximately unlike the other datasets.

The interestingness table attend balance value. It means there are two datasets in every method has the higher value and there is one data set which the approximately similar. The datasets by PARCD which are higher than MOPAR method are basket ball and bolt (6.56 and 43.43 %). Where the value of bolt is two time higher than opponent. On the other hand, the two remain datasets which are quake and body fat by MOPAR method are double than their value in PARCD. Then, the last dataset which is almost the same is pollution. It is about 10 % (Table 2.7).

Table 2.7 The comparison of coverage value

Dataset	Coverage (%)	
	PARCD	MOPAR
Quake	71.32	59.43
Basket ball	82.55	87.5
Body fat	99.48	6.28
Bolt	95.37	89
Pollution	91.58	89

The coverage table depicts that almost the datasets by PARCD method higher than MOPAR method which are the average over 90 % except quake dataset. The significantly highest gap value is body fat by 93 % and the lowest ones are pollution (just under 2 %). Then the remains are basket ball, pollution and quake the gaps are amount 5, 6 and 12 % respectively. Interestingly, the percentage of coverage of bolt and pollution by MOPAR method has the same value which are 89 %.

This table strongly give reason that the PARCD method can reach wider than the MOPAR method to searching the optimal value. This is because the proposed method, PARCD contain combination between PSO and Cauchy distribution which empirically prevent the PSO traps in local optima. It also makes additional evidence that this combination robust to solve some problems in different field including the numerical association rule mining optimization problem.

2.5 Conclusions

Based on this study, the weakness of PSO for solving numerical association rule mining problem can be solved by combining with Cauchy distribution. The problem of PSO that premature in minimum optima for searching in large dataset can be handled well in multi-objective function. The experiment by the method of PARCD showed obviously that in every multi-objective function such as confidence, comprehensibility and interestingness give the result better than previous method (MOPAR) which is only using PSO for solving multi-objective in numerical association rule mining problem.

For the future, because of the problem of numerical association rule mining is still be improved so it will be better to following the research for instance it combining with other methods like genetic algorithm or fuzzy algorithm.

Acknowledgments This research supported by various parties. We would like to thank for scholarship program from Kanazawa University, Japan and Ministry of Research, Technology and Higher Education (KEMENRISTEKDIKTI) and also STMIK AMIKOM Purwokerto, Indonesia. In addition, we thank for anonymous reviewers who gave input and correction for improving this research.

References

1. Arotaritei D, Negoita MG (2003) An optimization of data mining algorithms used in fuzzy association rules. In: Knowledge-based intelligent information and engineering systems. Springer, pp 980–985
2. Beiranvand V, Mobasher-Kashani M, Bakar AA (2014) Multi-objective PSO algorithm for mining numerical association rules without a priori discretization. *Expert Syst Appl* 41(9):4259–4273
3. Yan X, Zhang C, Zhang S (2009) Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. *Expert Syst Appl* 36(2):3066–3076
4. Almasi M, Abadeh MS (2015) Rare-pears: a new multi objective evolutionary algorithm to mine rare and non-redundant quantitative association rules. *Knowl Based Syst* 89:366–384
5. Minaei-Bidgoli B, Barmaki R, Nasiri M (2013) Mining numerical association rules via multi-objective genetic algorithms. *Inf Sci* 233:15–24
6. Li C, Liu Y et al (2007) A fast particle swarm optimization algorithm with cauchy mutation and natural selection strategy. In: Advances in computation and intelligence. Springer, pp 334–343
7. Alhajj R, Kaya M (2008) Multi-objective genetic algorithms based automated clustering for fuzzy association rules mining. *J Intell Inf Syst* 31(3):243–264
8. Alatas B, Akin E, Karci A (2008) Modenar: multi-objective differential evolution algorithm for mining numeric association rules. *Appl Soft Comput* 8(1):646–656
9. Kaya M (2006) Multi-objective genetic algorithm based approaches for mining optimized fuzzy association rules. *Soft Comput* 10(7):578–586
10. Qodmanan HR, Nasiri M, Minaei-Bidgoli B (2011) Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. *Expert Syst Appl* 38(1):288–298
11. Alatas B, Akin E (2008) Rough particle swarm optimization and its applications in data mining. *Soft Comput* 12(12):1205–1218
12. Indira K, Kanmani S (2015) Association rule mining through adaptive parameter control in particle swarm optimization. *Comput Stat* 30(1):251–277
13. Álvarez VP, Vázquez JM (2012) An evolutionary algorithm to discover quantitative association rules from huge databases without the need for an a priori discretization. *Expert Syst Appl* 39(1):585–593
14. Sarath K, Ravi V (2013) Association rule mining using binary particle swarm optimization. *Eng Appl Artif Intell* 26(8):1832–1840
15. Sangsawang C, Sethanan K et al (2015) Metaheuristics optimization approaches for two-stage reentrant flexible flow shop with blocking constraint. *Expert Syst Appl* 42(5):2395–2410
16. Yu M, Zhang Y et al (2015) Integration of process planning and scheduling using a hybrid GA/PSO algorithm. *Int J Adv Manuf Technol* 78(1–4):583–592
17. Gen M, Lin L, Owada H (2015) Hybrid evolutionary algorithms and data mining: case studies of clustering. In: Proceedings of social plant engineering Japan 2015 autumn conference
18. Ghosh A, Nath B (2004) Multi-objective rule mining using genetic algorithms. *Inf Sci* 163(1):123–133

Proceedings of the Tenth International Conference on
Management Science and Engineering Management

Xu, J.; Hajiyeu, A.; Nickel, S.; Gen, M. (Eds.)

2017, LI, 1723 p. 375 illus., 191 illus. in color. In 2
volumes, not available separately., Softcover

ISBN: 978-981-10-1836-7