International Journal of Power Control Signal and Computation(IJPCSC)
Vol3. No1. Jan-Mar 2012 ISSN: 0976-268X
www.ijcns.com

An Optimization of Association Rule Mining Algorithm using Weighted Quantum behaved PSO

S.Deepa[1], M. Kalimuthu[2]
[1]PG Student, Department of Information Technology
[2]Associate Professor, Department of Information Technology
SNS College of Technology
Coimbatore, Tamilnadu, India
[1]deepasengg@gmail.com
[2]mkmuthu73@gmail.com

Abstract

In this paper we propose Weighed Quantum behaved Particle Swarm Optimization (WQPSO) algorithm for improving the performance of association rule mining algorithm Apriori. It is a global convergence guaranteed algorithm, which outperforms original PSO algorithm and it has fewer parameters to control the search ability of PSO. Finding minimum support and minimum confidence values for mining association rules seriously affect the quality of association rule mining. In association rule mining, the minimum threshold values are always given by the user. But in this paper, WQPSO algorithm is used to determine suitable threshold values automatically and also it improves the computational efficiency of Apriori algorithm. First, the WQPSO algorithm is processed to find the minimum threshold values. In this algorithm which particle having the highest optimal fitness value, its support and confidence values are taken as the minimum threshold value to association rule algorithm. Then the minimum support and minimum confidence values are given to the input of Apriori association rule mining algorithm for mining association rules. Thus the proposed algorithm is verified by applying the FoodMart2000 database to Microsoft SQL Server 2000. The experimental results show that our proposed method gives better performance and less computational time than the existing algorithms.

***Index Terms*—Apriori algorithm, Association rule mining, Data mining, PSO algorithm.**

1. INTRODUCTION

Data mining is "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data." Data mining is an inter-disciplinary field, whose core is at the intersection of machine learning, statistics and databases. Data mining can be categorized into several models, including association rules, clustering and classification. Among these models, association rule mining is the most widely applied method. In the area of association rule mining, most previous research had focused on improving computational efficiency. The Apriori algorithm is the most representative algorithm. It consists of many modified algorithms that focus on improving its efficiency and accuracy. However, two parameters, minimal support and confidence, are always determined by the decision-maker him/herself or through trial-and-error, which seriously affect the quality of association rule mining, is still under investigation.

Particle swarm optimization (PSO), first introduced by Kennedy and Eberhart [3], is a population-based optimization technique, where a population is called a swarm. A simple explanation of the PSO's operation is as follows. Each particle represents a possible solution to the optimization task at hand. During each iteration, the accelerating direction of one particle determined by its own best solution found so far and the global best position discovered so far by any of the particles in the swarm. This means that if a particle discovers a promising new solution, all the other particles will move closer to it, exploring the region more thoroughly in the process. As far as the PSO itself concerned, however, it is not a global optimization algorithm, as has been demonstrated by Van den Bergh [4]. In [5,6], Sun et al. introduce quantum theory into PSO and propose a quantum-behaved PSO (QPSO) algorithm, which can be guaranteed theoretically to find optimal solution in search space. The experiment results on some widely used benchmark functions show that the QPSO works better than standard PSO and should be a promising algorithm. In this paper, in order to balance the global and local searching abilities, we introduce a weight parameter in calculating the mean best position in QPSO to render the importance of particles in population when they are evolving, and thus proposed an improved quantum-behaved particle swarm optimization algorithm, weighted QPSO (WQPSO).

The most representative association rule algorithm is the Apriori algorithm, which was proposed by Agrawal et al. in 1993. The Apriori algorithm repeatedly generates candidate itemsets and uses minimal support and minimal confidence to filter these candidate itemsets to find high-frequency itemsets. Association rules can be figured out from the high-frequency itemsets [2].

The rest part of the paper is organized as follows. In Section 2, a brief introduction of Association rule mining

algorithm is given. The QPSO and its related work is introduced in Section 3. In Section 4, we propose the of particles. Some experiments result on functions and discussions are presented in Section 5. Finally, the paper is concluded in Section 6.

2. ASSICIATION RULE MINING

This section briefly presents the general algorithms of association rule mining. Section A defines the basic definition of association rule mining, Section B defines the association rule mining algorithm Apriori and Section C defines the other association rule mining algorithms, which are used for mining association rules.

A. Definition

The Association rule mining defines that, some hidden relationships exist between purchased items in transactional databases. Therefore, mining results can help decision-makers understand customers' purchasing behavior. An association rule is in the form of X→Y, where X and Y represent Itemset(I), or products, respectively and Itemset includes all possible items. However, the mining association rule must accord with two parameters at the same time:

(1)Minimal support: Finding frequent itemsets with their supports above the minimal support threshold.

$$(\rightarrow) = \frac{\#}{\#} \quad \& \tag{1}$$

(2)Minimal confidence: Using frequent itemsets found in Eq. (1) to generate association rules that have confidence levels above the minimal confidence threshold.

$$(\rightarrow) = \frac{\#}{\#} \quad \& \tag{2}$$

B. Apriori Algorithm

The most representative association rule algorithm is the Apriori algorithm, which was proposed by Agrawal et al. in 1993. The Apriori algorithm repeatedly generates candidate itemsets and uses minimal support and minimal confidence to filter these candidate itemsets to find high-frequency itemsets. Association rules can be figured out from the high-frequency itemsets. The process of finding high-frequency itemsets from candidate itemsets is introduced in Fig. 1 [1].

In Fig. 1, Step 1.1 finds the frequent itemset, represented as L1. In Steps 1.2 through 1.10, L1 are utilized to generate candidate itemset Ck to find Lk. The process "Apriori gen" generates candidate itemsets and processes join and prune. The join procedure, Steps 2.1–2.4, combines Lk−1 into candidate itemsets. The prune procedure, Steps 2.5–2.7, deletes infrequent candidate itemsets. Infrequent itemset is tested in "has infrequent subset." After the Apriori algorithm has generated frequent itemsets, association rules can be generated. As long as the calculated confidence of a frequent itemset is larger than the predefined minimal confidence, its corresponding association rule can be accepted. Since the processing of the Apriori algorithm requires plenty of time, its computational efficiency is a very important issue. In order to improve the efficiency of Apriori, WQPSO algorithm is proposed.

improved WQPSO and show how to balance the searching abilities to guarantee the better convergence speed

```
Input : Database of transactions (D), Minimal Support threshold (min_sup)
Output : Frequent itemsets in D (L)
Method :
1.1  L1=find_frequent_1-itemsets(D);
1.2  for(k=2 ; L1-k ≠ φ ; k++){
1.3      Ck =apriori_gen( L1-k ,min_sup);
1.4      for each transaction t∈ D{     // scan D for counts
1.5          Ct =subset( Ck ,t);     //get the subsets of t that are candidates
1.6          for each candidate c∈ Ct
1.7              c.count++;
1.8      }
1.9      Lk ={c∈ Ck |c.count ≥ min_sup}
1.10 }
1.11 return L=∪k Lk ;
Procedure apriori_gen
2.1  for each itemset l1 ∈ L1-k
2.2  for each itemset l2 ∈ L1-k
2.3  if (l1[1] = l2[1]) ∧ (l1[2] = l2[2]) ∧ ...(l1[k -1] = l2[k -1]) then{
2.4  c= l1 ⊳⊲ l2 ;     //join step:generate candidates
2.5  if has_infreqrent_subset(c, L1-k )then
2.6  delete c;     //prune step:remove unfruitful candidate
2.7  else add c to Ck ;
2.8  }
2.9  return Ck ;
Procedure has_infrequent_subset
3.1  for each(k-1)-subset s of c
3.2  if s∉ L1-k then
3.3  return TRUE;
```

Figure. 1 The Apriori algorithm

C. Other Association rule mining algorithms

In this section defines, the other algorithms used for optimization of association rule mining.

An efficient hash-based method for discovering the maximal frequent set (HMFS) algorithm. In 2001, Yang et al. proposed the efficient hash-based method, HMFS, for discovering maximal frequent itemsets. The HMFS method combines the advantages of both the DHP and the Pincer-Search algorithm. The combination of the two methods leads to two advantages. First, the HMFS method, in general, can reduce the number of database scans. Second, the HMFS can filter the infrequent candidate itemsets and use the filtered itemsets to find the maximal frequent itemsets. These two advantages can reduce the overall computing time of finding the maximal frequent itemsets. In addition, the HMFS method also provides an efficient mechanism to construct the maximal frequent candidate itemsets so as to reduce the search space [7].

Genetic algorithms have also been applied in association rule mining [8]. This study uses weighted items to represent the importance of individual items. These weighted items are applied to the fitness function of heuristic genetic algorithms to estimate the value of different rules. These genetic algorithms can generate suitable threshold values for association rule mining. In addition, Saggar et al. proposed an approach concentrating on optimizing the rules generated using genetic algorithms. The most important aspect of their approach is that it can predict the rule that contains negative attributes [9]. According to the test results, the conclusion drawn stated that the genetic algorithm had considerably higher efficiency [10].

Particle swarm optimization algorithm Kennedy and Eberhart proposed the particle swarm optimization (PSO) algorithm in 1995. The PSO algorithm has become an evolutionary computation technique and an important heuristic algorithm in recent years. The main concept of PSO originates from the study of fauna behavior [12].

3. REALATED WORK

This related work defines the optimization algorithms such as PSO and our proposed algorithm WQPSO. Section A defines the Particle Swarm Optimization algorithm and its velocity, position update. Section B defines that quantum behaved Particle Swarm Optimization algorithm. At last Section C defines our proposed Weighted Quantum Behaved PSO.

A. Particle Swarm Optimization (PSO) algorithm

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. The fitness value is also stored. 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 with Eqs. (3) and (4), as follows [11]:

Velocity calculation is defined as,

$$v_{id(t)} = \omega v_{id(t-1)} + c_1 \times rand() \times (p_{id} - x_{id}) + c_2 \times Rand() \times (p_{gd} - x_{id}) \quad (3)$$

Position update is defined as,

$$x_{id(t)} = x_{id(t-1)} + v_{id(t)} \quad (4)$$

In the above equation, x_{id} is a current value of the dimension "d" of the individual "i", v_{id} is a current velocity of the dimension "d" of the individual "i", P_{id} is a optimal value of the dimension "d" of the individual "i" so far, P_{gd} is a current optimal value of the dimension "d" of the swarm, c_1, c_2 are a acceleration coefficients, w is a inertia weight factor. This PSO algorithm is not a global convergence algorithm and which not having any control of its search ability. So WQPSO algorithm is proposed for global convergence and for real time applications.

B. Quantum behaved PSO

In practice, the evolution of mans thinking is uncertain to a great extent somewhat like a particle having quantum behavior. In [3], Jun Sun *er al*. introduce quantum theory into PSO and propose a Quantum behaved PSO (QPSO) algorithm Their experiment results indicate that the QPSO works better than standard PSO on several benchmark functions and it is a promising algorithm. In this section a novel parameter control method of QPSO *is* defined. The new approach in the revised QPSO, with a global reference point called "Mainstream Thought" introduced to evaluate the search scope of a particle, is more efficient in global search than that in [3].

In quantum-behaved PSO, search space and solution space of the problem are two spaces of different quality. Wave function or probability function of position depicts the state of the particle in quantized search space, not informing us of any certain information about the position of a particle that is vital to evaluate the fitness of a particle. Therefore, state transformation between two spaces is absolutely necessary In terms of quantum mechanics, the transformation from quantum state to classical state is called collapse, which in nature is the measurement of a particle s position. In a quantized search space, wave function is defined for finding global search solution

C. Weighted quantum behaved PSO

In this section, we propose an improved quantum-behaved particle swarm optimization with weighted mean best position according to fitness values of the particles. It is shown that the improved QPSO has faster local convergence speed, resulting in better balance between the global and local searching of the algorithm, and thus generating good performance. In summary, the main contributions of this paper are as follows:

1. We propose the specific algorithm to determine the minimum threshold values using Weighted Quantum behaved Particle Swarm Optimization algorithm, and which gives reliable and efficient values.
2. Then we apply the minimum threshold values to Apriori association rule mining algorithm, to improve the performance and efficiency of association rules for real time transactions.

4. WEIGHTED QUANTUM BEHAVED PARTICLE SWARM OPTIMIZATION ALGORITHM

In this paper, in order to balance the global and local searching abilities, we introduce a weight parameter in calculating the mean best position in QPSO to render the importance of particles in population when they are evolving, and thus proposed an improved quantum-behaved particle swarm optimization algorithm, weighted QPSO (WQPSO). The proposed WQPSO algorithm comprises two parts, preprocessing and mining. The first part provides procedures related to calculating the fitness values of the particle swarm. Thus, the data are transformed and stored in a binary format. Then, the search range of the particle swarm is set using the IR (itemset range) value. In the second part of the algorithm, which is the main contribution of this study, the WQPSO algorithm is employed to mine the association rules.

A. Binary transformation

This study adopts the approach proposed by Wur and Leu in 1998 [13] to transform transaction data into binary type data, each recorded and stored as either 0 or 1. This approach can accelerate the database scanning operation, and it calculates support and confidence more easily and quickly.

B. IR Value calculation

This study applies the WQPSO algorithm in association rule discovery, as well as in the calculation of IR value which is included in chromosome encoding. The purpose of such an inclusion is to produce more meaningful association rules. Moreover, search efficiency is increased when IR analysis is

utilized to decide the rule length generated by chromosomes in particle swarm evolution. IR analysis avoids searching for too many association rules, which are meaningless itemsets in the process of particle swarm evolution. This method addresses the front and back partition points of each chromosome, and the range decided by these two points is called the IR, which is shown in Eq. (5):

$$IR = [\log(m\text{TransNum}(m)) + \log(n\text{TransNum}(n))]\frac{\text{Trans}(m, n)}{\text{TotalTrans}} \quad (5)$$

In Eq. (5), m / = n and m< n. "m" represents the length of the itemset and TransNum(m) means the number of transaction records containing m products. "n" is the length of the itemset, and TransNum(n) means the number of transaction records containing n products. Trans(m, n) means the number of transaction records purchasing m to n products. Total Trans represents the number of total transactions.

The WQPSO algorithmic process is quite similar to that of PSO algorithms, but the proposed procedures include only the main Quantum function and its mean best position. Each of the steps in the WQPSO algorithm and the process of generating association rules are explained as follows:

C. Encoding

According to the definition of association rule mining, the intersection of the association rule of itemset X to itemset Y (X→Y) must be empty. Items which appear in itemset X do not appear in itemset Y, and vice versa. Hence, both the front and back partition points must be given for the purpose of chromosome encoding. The itemset before the front partition point is called "itemset X," while that between the front partition and back partition points is called "itemset Y." The chromosome encoding approach in this study is "string encoding."

D. Fitness value calculation

The fitness value in this study is utilized to evaluate the importance of each particle. The fitness value of each particle comes from the fitness function. Here, we employ the target function proposed by Kung [14] to determine the fitness function value. The improved algorithm is called Weighted QPSO that is outlined as follows.

$$\text{Fitness}(k) = \text{confidence}(k)X\log(\text{support}(k)X\,\text{length}(k) + 1) \quad (6)$$

Fitness (k) is the fitness value of association rule type k. Confidence (k) is the confidence of association rule type k. Support (k) is the actual support of association rule type k. Length (k) is the length of association rule type k. The objective of this fitness function is maximization. The larger the particle support and confidence, the greater the strength of the association, meaning that it is an important association rule.

E. Population generation:

In order to apply the evolution process of the WQPSO algorithm, it is necessary to first generate the initial population. In this study, we select particles which have larger fitness values as the population. The particles in this population are called initial particles.

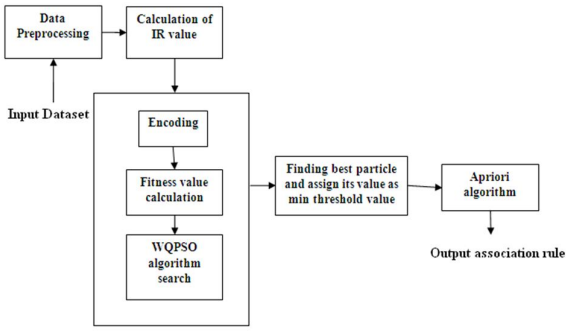

Figure. 2 The Proposed system architecture

F. Search the best particle

First, the particle with the maximum fitness value in the population is selected as the "gbest".

From Eq(7), we can see that the mean best position is simply the average on the personal best position of all particles, which means that each particle is considered equal and exert the same influence on the value of m. The philosophy of this method is that the Mainstream Though, that is, mean best position m, determines the search scope or creativity of the particle [14]. The definition of the Mainstream Thought as mean of the personal best positions is somewhat reasonable.

$$m(t) = (m_1(t), m_2(t), \ldots, m_3(t)) = \left(\frac{1}{M}\sum_{i=1}^{M} P_{i,1}(t), \frac{1}{M}\sum_{i=1}^{M} P_{i,2}(t), \ldots, \frac{1}{M}\sum_{i=1}^{M} P_{i,n}(t)\right), \quad (7)$$

The equally weighted mean position, however, is something of paradox, compared with the evolution of social culture in real world. For one thing, although the whole social organism determines the Mainstream Thought, it is not properly to consider each member equal. In fact, the elitists play more important role in culture development. With this in mind when we design a new control method for the QPSO in this paper, m in Equation of QPSO is replaced for a weighted mean best position. The most important problem is to determine whether a particle is an elitist or not, or say it exactly, how to evaluate its importance in calculate the value of m. It is natural, as in other evolutionary algorithm, that we associate elitism with the particles' fitness value. The greater the fitness, the more important the particle is. Describing it formally, we can rank the particle in descendent order according to their fitness value first. Then assign each particle a weight coefficient ai linearly decreasing with the particle's rank, that is, the nearer the best solution, the larger its weight coefficient is. The mean best position m, therefore, is calculated as

$$m(t) = (m_1(t), m_2(t), \ldots, m_3(t)) = \left(\frac{1}{M}\sum_{i=1}^{M} a_{i,1}P_{i,1}(t), \frac{1}{M}\sum_{i=1}^{M} a_{i,2}P_{i,2}(t), \ldots, \frac{1}{M}\sum_{i=1}^{M} a_{i,n}P_{i,n}(t)\right) \quad (8)$$

G. Termination condition:

To complete particle evolution, the design of a termination condition is necessary. In this study, the evolution terminates when the fitness values of all particles are the same. In other words, the positions of all particles are fixed. Another termination condition occurs after 100 iterations and the

evolution of the particle swarm is completed. Finally, after the best particle is found, its support and confidence are recommended as the value of minimal support and minimal confidence. These parameters are employed for association rule mining to extract valuable information.

The WQPSO is much different from the PSO in that the update equation of QPSO ensures the particle's appearing in the whole n-dimensional search space at all iterations, while the particle in the PSO can only fly in a bounded space at each iteration. Employing the global convergence criterion, we can conclude that the QPSO or WQPSO is a global convergent algorithm and the PSO is not.

5. EXPERIMENTAL STUDIES

Our proposed methodology is to be evaluated for the Food mart 2000 database using Microsoft SQL server and java. This experiment also compares the performance of existing PSO algorithm and the proposed WQPSO algorithm. This results shows that, WQPSO algorithm minimizes the Quantization errors and also it gives improved fitness values for mining association rules. By using this optimization algorithm, the performance of Apriori association rule mining algorithm will be improved.

The performance of this algorithm is measured using the following parameters. The parameters are no of iterations, quantization error, fitness value and accuracy of association rules. The quantization error is minimized using our proposed algorithm. The accuracy and fitness values are also to be improved.

A. Results and Discussion

Our proposed algorithm minimizes quantization errors and it gives large fitness value for all iterations. The results are discussed and compared with Particle swarm optimization algorithm.

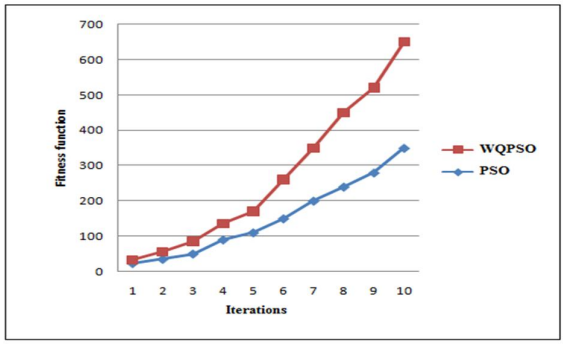

Figure. 3 Iterations Vs Fitness value

Figure 3 defines the iterations Vs fitness value. That is the fitness value is increased for WQPSO than the PSO algorithm. Thus the accuracy of association rules is also to be improved.

The average running times for different population sizes are also illustrated. Furthermore, in regard to the selection of threshold value setup, this study can provide the most feasible minimal support and confidence. This dramatically decreases the time consumed by trial-and-error. This algorithm mainly defined for improving the performance of the Apriori algorithm.

Figure 4 defines the iterations Vs quantization error. It defines that the quantization error of WQPSO is minimized using our proposed algorithm.

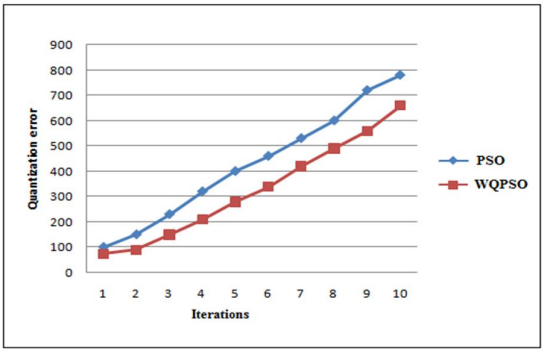

Figure. 4 Iterations Vs Quantization error

Thus, the proposed WQPSO algorithm is better than the traditional Apriori algorithm since it does not need to subjectively set up the threshold values for minimal support and confidence. This can also save computation time and enhance performance.

6. CONCLUSION

In the field of association rule mining, the minimum threshold values are always given by the user. But this study intends to determine the minimum support and minimum confidence values for mining association rules using the WQPSO optimization algorithm. This algorithm mainly defined for improving the performance of the Apriori algorithm. Thus also it minimizes the quantization errors and fitness value to be improved. From this experiment we can know that, WQPSO algorithm is a global convergence algorithm and PSO is not. Thus the Optimizing association rule mining algorithms gives better results than the simple association rule mining algorithms.

REFERENCES

[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, New York, 2000.

[2] A. Savasere, E. Omiecinski, S. Navathe, An efficient algorithm for mining association rules in large database, in: Proceedings of the 21st VLDB Conference, 1995, pp. 432–444.

[3] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of IEEE International Conference On Neural Network, 1995, pp. 1942–1948.

[4] F. Van den Bergh, An Analysis of Particle Swarm Optimizers, Ph.D. Thesis, University of Pretoria, November 2001.

[5] J. Sun, B. Feng, W.B. Xu, Particle swarm optimization with particles having quantum behavior, in: IEEE Proceedings of Congress on Evolutionary Computation, 2004, pp. 325–331.

[6] J. Sun, W.B. Xu, B. Feng, a global search strategy of quantum-behaved particle swarm optimization, in: Cybernetics and Intelligent Systems Proceedings of the 2004 IEEE Conference, 2004, pp. 111–116.

[7] D.L. Yang, C.T. Pan, Y.C. Chung, An efficient hash-based method for discovering the maximal frequent set, in: Proceeding of the 25th Annual International Conference on Computer Software and Applications, 2001, pp. 516–551.
[8] S.S. Gun, Application of genetic algorithm and weighted itemset for association rule mining, Master Thesis, Department of Industrial Engineering and Management, Yuan-Chi University, 2002.
[9] M. Saggar, A.K. Agrawal, A. Lad, Optimization of association rule mining using improved genetic algorithms, in: Proceeding of the IEEE International Conference on Systems Man and Cybernetics, vol. 4, 2004, pp. 3725–3729.
[10] C. Li, M. Yang, Association rule data mining in manufacturing information system based on genetic algorithms, in: Proceeding of the 3rd International Conference on Computational Electromagnetics and Its Applications, 2004, pp. 153–156.
[11] Particle Swarm Optimization: Tutorial, http://www.swarmintelligence.org/
tutorials.php
[13] M.P. Song, G.C. Gu, Research on particle swarm optimization: a review, in: Proceedings of the IEEE International Conference on Machine Learning and Cybernetics, 2004, pp. 2236–2241.
[14] C.Y. Chen, F. Ye, Particle swarm optimization algorithm and its application to clustering analysis, in: Proceedings of the IEEE International Conference on Networking, Sensing and Control, Taipei, Taiwan, 2004, pp. 21–23.
[15] R.J. Kuo, M.J. Wang, T.W. Huang, Application of clustering analysis to reduce SMT setup time—a case study on an industrial PC manufacturer in Taiwan, in: Proceedings of International Conference on Enterprise Information Systems, Milan, Italy, 2008.
[16] R.J. Kuo, F.J. Lin, Application of particle swarm optimization to reduce SMT setup time for industrial PC manufacturer in Taiwan, International Journal of Innovative Computing, Information, and Control, in press.