International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.1, No.7, November 2012 DOI:10.15693/ijaist/2012.v1i7.26-32

Association Rule Mining by Dynamic Neighborhood Selection in Particle Swarm Optimization

K Indira
Research Scholar/Department of CSE
Pondicherry Engineering College
Puducherry, India

S Kanmani
Professor/Department of IT
Pondicherry Engineering College
Puducherry, India

***Abstract*— Association Rule (AR) mining is one of the most studied tasks in data mining community with focus on improving computational efficiency. The standard Particle Swarm Optimization (PSO) is an evolutionary algorithm originally developed to simulate the behavior of birds and successfully applied for mining association rules. The problem with Particle swarm optimization algorithm is its trapping into local optima. This result in premature convergence of the algorithm affecting the efficiency of the rules mined. To improve the performance of PSO and maintain diversity of particles, a dynamic neighborhood selection in PSO is proposed for mining ARs. Dynamic neighborhood selection in PSO introduces the concept of local best particle (lBest) replacing the particle best (pbest). The algorithm when tested generates association rules with better predictive accuracy.**

Index terms - Association rules, Particle Swarm Optimization, Dynamic neighborhood selection PSO, Local best, Predictive Accuracy.

I. INTRODUCTION

Data mining is extracting nontrivial, implicit, previously unknown and potential information from large databases. Association rules, Clustering and Classification are methods applied for extracting information from databases. Association rule mining is the most widely applied method. Association rule mining to find interesting patterns or relation among data in large databases.

The Apriori algorithm is the most standard method for mining association rules. The efficiency and accuracy of the system mainly relies on the two parameters: the minimum support and minimum confidence. The methodology involves traversing the dataset many times, increasing the computational complexity and Input/output overhead. Traditional rule generation methods, are usually accurate, but have brittle operations. Evolutionary algorithms on the other hand provide a robust and efficient approach to explore large search space.

Evolutionary Algorithms (EAs) inspired by Darwinian's theory of biological evolution and natural selection have been popular search algorithm over recent years and inspired many research efforts for optimization as well as rule generation [7,8]. Particle Swarm optimization is an evolutionary computational technology proposed in 1995 by Kennedy and Eberhart where individuals interact with one another while learning from their own experience. The control parameters of PSO have significant effect on the performance of the algorithm.

The balance between exploration and exploitation in PSO is the main issue when applied for solving complex problem. To maintain the diversity of the particles and enhance the performance of the PSO, the concept of adapting the local best particle from the neighborhood is proposed for mining association rules. The proposed work is to review the PSO for mining association rules with dynamic neighborhood selection.

The remaining of the paper is organized as follows. Section 2 briefly presents the general background. The proposed method is explained in Section 3, while the computational results of the methodology on three datasets from UCI repository is presented in section 4. The concluding remarks are made in section 5.

II. Literature Review

This section presents the general background of association rule mining and particle swarm optimization, followed by related works on this subject.

A. Association Rule

Association rule mining [9] is the most broadly discussed area in data mining. The idea of mining association rules originates from the analysis of data from a market basket. A person buying goods x1 and x2 could also buy another product with probability c%. Association rules express how important products or services relate to each other, and immediately suggest particular actions. Association rules are used in mining categorical data – items.

In general, the association rule [10] is an expression of the form X=>Y, where X is antecedent and Y is consequent. Association rule shows how many times Y has occurred if X has already occurred depending on the support and confidence value. Each association rule has two quality measurements, support and confidence, defined as follows:
Support: The support indicates how often the rule holds in a set of data. This is a relative measure determined by dividing the number of data that the rule covers, i.e., total number of data that support the rule, by the total number of data in the

set. It is the probability of item or item sets in the given transactional data base:

$$\sup(x) = \frac{No.of\ transactions\ \ containing\ \ X}{Total\ \ No.of transactions} \quad (1)$$

where n is the total number of transactions in the database and n(X) is the number of transactions that contains the item set X.

Confidence: The confidence for a given rule is a measure of how often the consequent is true, given that the antecedent is true. If the consequent is false while the antecedent is true, then the rule is also false. If the antecedent is not matched by a given data item, then this item does not contribute to the determination of the confidence of the rule. It is conditional probability, for an association rule X →Y and defined as

$$conf\ (X \rightarrow Y) = \frac{\sup(X \cup Y)}{\sup(X)} \quad (2)$$

Mining association rules can be broken down into the following two sub-problems[11]:

- Generating all itemsets that have support greater than, or equal to, the user specified minimal support. That is, generating all large itemsets
- Generating all the rules that have minimum confidence

B. Particle Swarm Optimization

Particle Swarm Optimization is an innovative intelligent paradigm for solving optimization problems that incorporates swarming behaviors observed in flocks of birds, schools of fish, or swarms of bees, and even human social behavior, from which the idea is emerged [12,13]. As an algorithm, the main strength of PSO is its fast convergence, which compares favorably with many global optimization algorithms like Genetic Algorithms.

PSO is initialized with a group of random particles (solutions) and then searches for optima by updating generations. In every iteration, each particle is updated by following two "best" values. The first one is the best solution (fitness) it has achieved so far. This value is called pbest. Another "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 called gBest.
PSO has been introduced for classification rules mining in [14]. However it cannot cope directly with nominal attributes, that nominal values are converted into binary numbers in a preprocessing phase.

After finding the two best values, the velocity and position of each particle is updated with equations (3) and (4), as follows [15]:

$$v_{id}^{new} = v_{id}^{old} + c_1 rand()(pbest - x_{id}) + c_2 rand()(gbest - x_{id}) \quad (3)$$

$$x_{id}^{new} = x_{id}^{old} + v_{id}^{new} \quad (4)$$

Where
v_{id} is the particle velocity of the idth particle;
x_{id} is the idth, or current, particle;
i is the particle's number;
d is the dimension of searching space.
rand ( ) is a random number in (0, 1);
c_1 is the individual factor;
c_2 is the societal factor;
pbest is the particle best;
gBest is the global best.

Usually c1 and c2 are set to be 2 [16].

All particles have fitness values calculated by the fitness function. Particles velocities on each dimension are clamped to a maximum velocity V_{max}. If the sum of accelerations causes the velocity on that dimension to exceed V_{max}, which is a parameter specified by the user, then the velocity on that dimension is limited to V_{max}. This method is called V_{max} method [16]. Constriction factor [18] is mainly based on the individual factor c_1 and societal factor c_2 balances the effect of exploration and exploitation in velocity update function.

The outline of basic particle swarm optimizer is as follows
Step1. Initialize the population - locations and velocities
Step 2. Evaluate the fitness of the individual particle (pBest)
Step 3. Keep track of the individuals highest fitness (gBest)
Step 4. Modify velocities based on pBest and gBest position
Step 5. Update the particles position
Step 6. Terminate if the condition is met
Step 7. Go to Step 2

C. *Related Works*

PSO is a population-based, stochastic optimization algorithm based on the idea of a swarm moving over a given landscape. The algorithm adaptively updates the velocities and positions of the members of the swarm by learning from the good experiences.

The velocity update equation plays a major role in enhancing the performance of the PSO. To balance the global search and local search inertia weight (w) was introduced. It can be a positive constant or even a positive linear or nonlinear function of time [17]. In Gregarious PSO the social knowledge of the particle is used for discovery in the search space. If particles are trapped in the local optimum, a stochastic velocity vector thereby self sets the parameters [18]. In Dynamic neighborhood PSO [19] instead of using the current GBest, another parameter Nbest is utilized. This term is the

best particle among the current particle's neighbors in a specified neighborhood.

The fixing up of the best position [1] for particles after velocity updation by using Euclidean distance helps in generating the best particles. The chaotic operator based on Zaslavskii maps when used in velocity update equation [2] proved to enhance the efficiency of the method. The soft adaptive particle swarm optimization algorithm [3] exploits the self adaptation in improving the ability of PSO to overcome optimization problems with high dimensionality. The particle swarm optimization with self adaptive learning [4] aims in providing the user a tool for various optimization problems.

The problem of getting struck at local optimum and hence premature convergence is overcome by self adaptive PSO [5] where the diversity of population is maintained. This copes up with the deception of multiple local optima and reduces computational complexity. An adaptive chaotic particle swarm optimization (cPSO) [6] enhances the global searching capability and local searching capability by introducing chaotic operators based on Logistic map and tent map. In addition novel adaptive search strategy which optimizes continuous parameters is employed.

III. Methodology

Genetic algorithm has been the active research focus for mining association rules recently. The accuracy of the association rules mined using genetic algorithm has proved to be enhanced when compared with existing standard methods [22, 23, 24]. The drawback of the GA is that it does not assure constant optimization results. GAs have the tendency to converge towards local optima or even arbitrary points rather than the global optimum of the problem and genetic algorithms do not scale well with complexity. That is, where the number of elements which are exposed to mutation is large there is often an exponential increase in search space size.

Particle swarm optimization with its velocity update and position update activities instead of the mutation and crossover operators (reproduction) of genetic algorithm avoids the problem of inconsistency in optimization over runs and the complexity of the algorithm is also simplified. The particle best (pbest) and global best (gBest) values tends to avoid premature convergence during optimization.

The problem of deviation from optimal solution space (exploitation) and not reaching the optimal solution in roundabout way (exploration) are addressed via gBest and pbest values respectively in particle swarm optimization. The global best propagates information the fastest in the population dealing with exploration; while the local best using a ring structure speeds up the system balancing the exploration.

Particle swarm optimization when applied for mining association rules results in earlier convergence than genetic algorithms. So to avoid premature convergence and enhance the accuracy the neighborhood selection in PSO was introduced replacing the particle best concept by local best.

Based on the background presented in the above section, this section proposes an algorithm, which applies particle swarm optimization with dynamic neighborhood selection in generating association rules from database.

A. Proposed Algorithm

The initial population is selected based on fitness value. The velocity and position of all the particles are set randomly. Based on the fitness function the importance of the particles is evaluated. The fitness function designed is based on support and confidence of the association rule. The objective of fitness function is maximization. The fitness function is shown in equation 5.

$$Fitness(x) = conf(x) \times \log(\sup(x) \times length\,(x) + 1) \quad (5)$$

Fitness (k) is the fitness value of association rule type k. conf (x) is the confidence of association rule type k. sup(x) is the actual support of association rule type k. When the support and confidence values are larger, then larger is the fitness value meaning that it is an important association rule.

The flowchart of the proposed algorithm is given in figure1.The particle with maximum fitness is selected as the 'gBest' particle. Each initial particle is considered as its 'lBest'. With velocity updation in each generation the gBest and lBest are updated. The neighborhood best (lBest) selection is as follows;

- Calculate the distance of the current particle from other particles by equation 6.

$$\Delta x_i = |(x_i - x_{gbest})| \quad (6)$$

- Find the nearest m particles as the neighbor of the current particle based on distance calculated
- Choose the local optimum lBest among the neighborhood in terms of fitness values

The number of neighborhood particles m is set to 2. Velocity and position updation of particles are based on equation 3 and 4. The velocity updation is restricted to maximum velocity V_{max} set by the user. The termination condition is set as fixed number of generations.

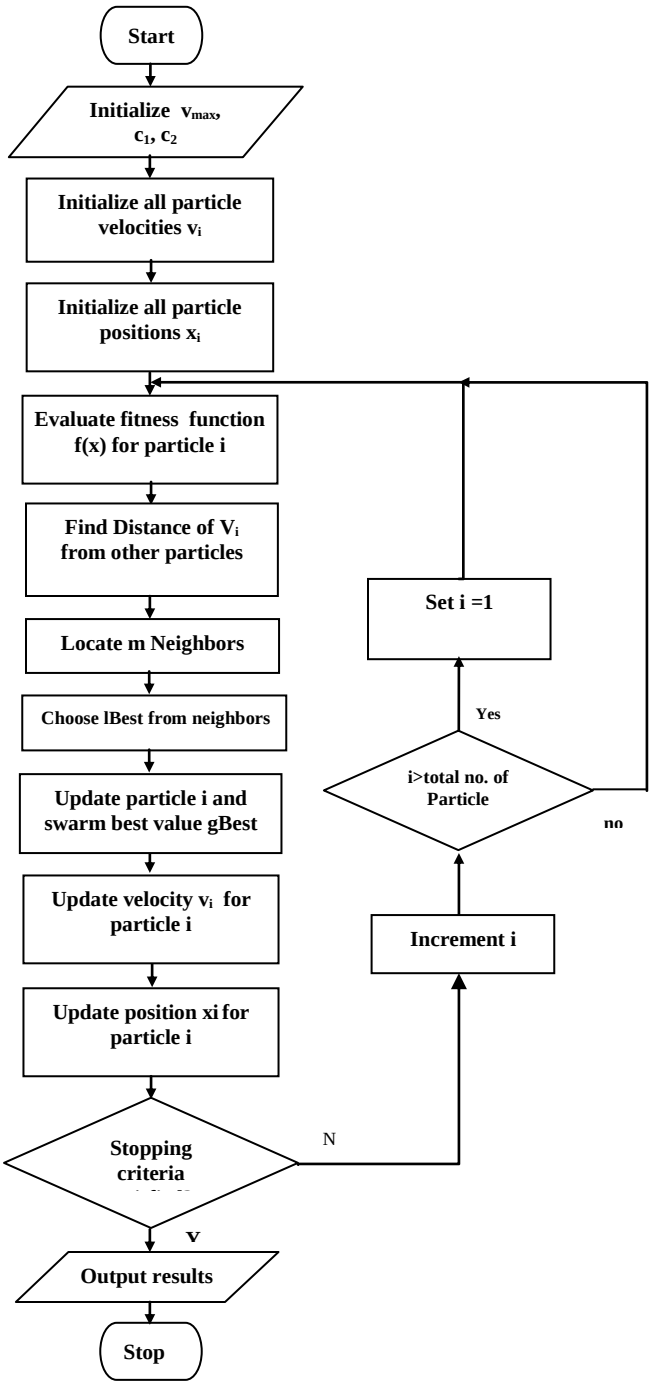


Fig 1. PSO algorithm for association rule mining with Dynamic Neighborhood Selection.

B. Predictive Accuracy

Predictive accuracy measures the effectiveness of the rules mined. The mined rules must have high predictive accuracy.

$$Predictive\ accuracy = \frac{|X\ \&\ Y|}{|X|} \quad (7)$$

where |X&Y| is the number of records that satisfy both the antecedent X and consequent Y, |X| is the number of rules satisfying the antecedent X.

C. Interestingness Measure

The process of discovering interesting and unexpected rules from large data sets is known as association rule mining. The interestingness of discovered association rules is an important and active area within data mining research. The measure of interestingness varies from application to application and from expert to expert. Each interestingness measure produces different results, and experts have different opinions of what constitutes a good rule. The interestingness measure for a rule is taken from relative confidence and is as follows:

$$interestingmess(k) = \frac{\sup(x \cup y) - \sup(x)\sup(y)}{\sup(x)(1-\sup(y))} \quad (8)$$

Where k is the rule, x the antecedent part of the rule and y the consequent part of the rule k.

IV. Evaluation Results and Discussion

Three datasets from University of California Irvine Machine Learning Repository namely Car Evaluation, Haberman's Survival and Lenses are taken up for evaluating the PSO with dynamic neighborhood selection algorithm.

Car evaluation dataset contains 1728 instances of records with 6 attributes and the swarm size set was 700. The Haberman's Survival dataset contains 306 instances of records with 3 attributes and the swarm size set was 300 and the Lenses dataset contains 24 instances of records with 3 attributes and the swarm size set was 24. The experiment was conducted on Microsoft windows XP platform using Java as the developing environment. Maximum number of iterations carried out was 50. The parameters set are given in table 1.

Table 1. Initial values set for the control Parameters

Parameter Name	Value for Lens	Value for Car Evaluation	Value for Haberman's Survival
Population Size	15	300	100
Initial Velocity	0	0	0
C_1	2	2	2
C_2	2	2	2
V_{max}	1	1	1

The maximum accuracy achieved from repeated runs is recorded as the predictive accuracy for each case. The interestingness is calculated from the corresponding run.

The predictive accuracy achieved is compared with PSO and Self Adaptive Genetic Algorithm (SAGA) [21] for the same datasets. The highest predictive accuracy achieved for multiple runs is plotted in figure 2.

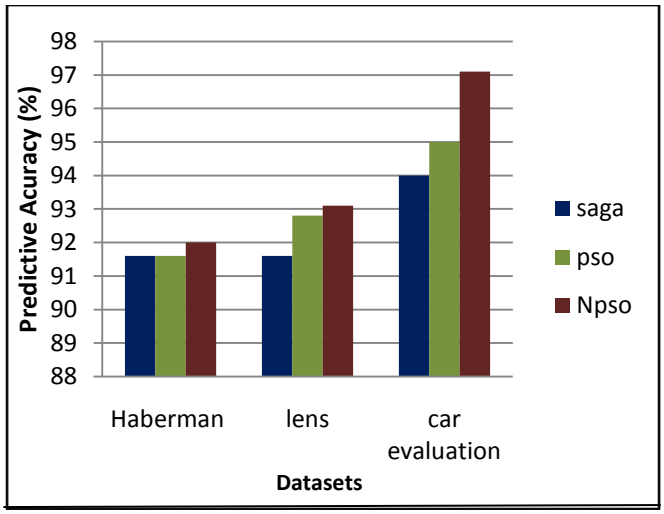


Figure 2. Predictive Accuracy Comparison for Dynamic Neighborhood selection in PSO

All three datasets produced enhanced results over SAGA and PSO methods. Both lenses and haberman's survival having less dataset shows marginal increase in accuracy, whereas car evaluation dataset for which the size and number of attributes is more the accuracy enhancement is noticeable.

The interestingness or relative confidence of the mined rule for the predictive accuracy is plotted in figure 2 is shown in table 2.

Table 2. Measure of Interestingness for Dynamic neighborhood selection PSO

Dataset	Interestingness Value
Lens	0.82
Car Evaluation	0.73
Haberman's Survival	0.8

The association rules mined with dynamic neighborhood selection PSO is with good interestingness measure indicating the importance of the mined rules.

Experts using evolutionary algorithms observe that the time complexity of particle swarm optimization is less when compared with genetic algorithm. Premature convergence may also result in reduced execution time. The scope of this work is to avoid premature convergence. The concept of local best based on neighborhood particles rather than individual particles focus on this.

The fixation of local best depends on search over neighborhood rather than individual particles best, increasing the search time and hence increase in execution time marginally. This is shown in figure 3 for all the three datasets in comparison with PSO.

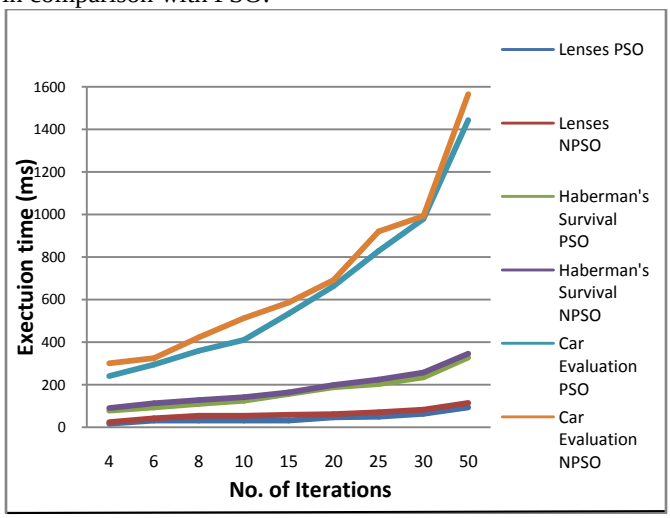
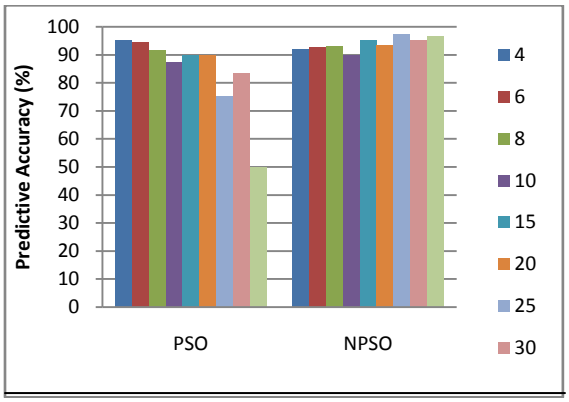
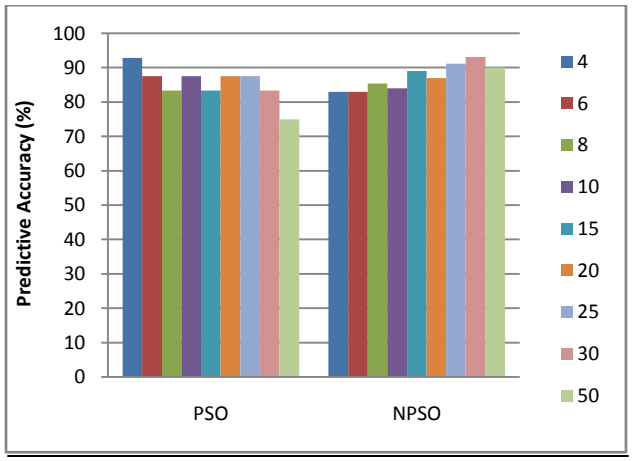


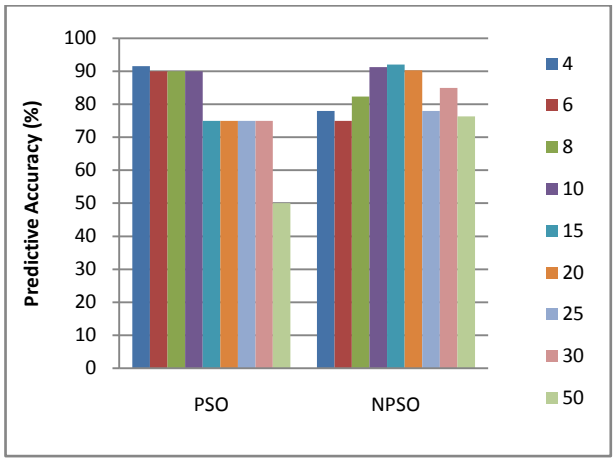
Figure 3. Execution Time Comparison for Dynamic Neighborhood selection in PSO



(a)



(b)



(c)

Figure 4. Predictive Accuracy over Generation for a) Car Evaluation b) Lenses c) Haberman's Survival datasets

The predictive accuracy over generations recorded for the three datasets is shown in figure 4. The predictive accuracy of lenses and car evaluation dataset are at optimum only at later generation whereas for haberman's survival dataset the optimal accuracy is achieved at earlier generations but better than PSO. The neighborhood selection in PSO extends the convergence rate avoiding premature convergence.

The dynamic neighborhood selection PSO mines association rules with better predictive accuracy when compared with PSO and SAGA with only marginal difference in execution time. The avoidance of premature convergence at local optimal points tends to enhance the results when compared with other methods.

The selection of local best particles based on neighbors (lBest) rather than particles own best (pbest) enhances the accuracy of the rules mined. The concept of local best (lBest) based on neighborhood selection in fitness space instead of other measures helps in maintaining the diversity of local points optimally, balancing between premature convergence and diversification of particles in problem space.

V. Conclusion

Particle swarm optimization is a recent heuristic search method based on the idea of collaborative behavior and swarming in populations. The problem with PSO is the balancing between exploration and exploitation of particles in problem space. The dynamic neighborhood

Selection PSO avoids premature convergence at local optima. The local best (lBest) based on neighborhood instead of the particle best (pBest) maintains the diversification of pbest positions and balances the premature convergence. The association rues mined has better accuracy when compared to PSO. The measure of interestingness of the mined rules is also good.

The neighborhood selection is done in fitness space rather than particles position from one another. The number of neighborhoods can be increased and the problem of exploration and exploitation by balancing the global best particle could be taken up for further study on the topic.

REFERENCES

1. Kuo R.J, Chao C.M, Chiu Y.T, "Application of Particle Swarm optimization in association rule mining", Applied soft computing, 11,2011, pp.323-336.
2. Bilal Atlas, Erhan Akin, "Multi-objective rule mining using a chaotic particle swarm optimization algorithms, Knowledge based systems", 2009, 23, pp. 455-460.
3. Yamina Mohammed Ben Ali, "Soft Adaptive particle swarm algorithm for large scale optimization", IEEE fifth international conference on bio inspired computing, 2010, pp.1658-1662.
4. Yu Wang, Bin LI, Thomas Weise, Jianyu Wang, Bo Yun, Qiondjie Tian, "Self-adaptive learning based on particle swarm optimization, Information Science", 2011, 181, pp.4515-4538.
5. Feng Lu, Yanfen Ge, Liqun Gao, "Self Adaptive particle swarm optimization algorithm for global optimization", Sixth IEEE international conference on natural computation,2010, pp.2692-2696.
6. Weijian Cheng, Jinliang Ding, Weijian Kong, Tianyou Chai, and S.Joe Qin, "An Adaptive Chaotic PSO for Parameter Optimization and Feature Extraction of LS-SVM Based Modelling", American Control Conference, 2011
7. A.A. Freitas, "Data Mining and Knowledge Discovery with Evolutionary Algorithms", Springer-Verlag, New York, 2002.
8. C.M. Fonseca, P.J. Fleming, "An overview of evolutionary algorithmsin multi-objective optimization", Evolutionary Computation, 1995, 3 (1),pp.1–16.
9. Agrawal, R., Imielinski, T., & Swami, A, "Mining association rules between sets of items in large databases", In Proceedings of ACM SIGMOD conference on management of data ,1995, pp. 207–216.
10. Berry, Linoff Michael J.A. Berry, Gordon Linoff. "Data Mining Techniques". John Wiley & Sons, 1997.
11. Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques," Multiscience Press, 1999.
12. Parsopoulos, K. E., and Vrahatis, M. N, "On the computation of all global minimizers through particle swarm optimization", IEEE Transactions on Evolutionary Computation,2004, 8(3), pp.:211-224.
13. Kennedy, J., and Eberhart, R. "*Swarm intelligence",* Morgan Kaufmann Publishers,Inc., San Francisco, CA. 2001
14. Sousa, T., Silva, A., Neves, "A. Particle Swarm based Data Mining Algorithms for classification tasks", *Parallel Computing 2004, 30,pp.*767–783
15. Particle Swarm Optimization: Tutorial, http://www.swarmintelligence.org/tutorials.php.

16. 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.
17. Y. Shi, R.C. Eberhart, "A modified particle swarm optimizer", in: Proceedings of the IEEE International Conference on Evolutionary Computation, Piscataway, 1998, pp. 69–73.
18. Pasupuleti, S. and Battiti, R.,. "The Gregarious Particle Swarm Optimizer (GPSO)", GECCO'06.
19. Lu, H. and Chen, W. , "Self-adaptive velocity particle swarm optimization for solving constrained optimization problems", J. of Global Optimization, 2008, Vol.41, No.3, pp. 427-445.
20. Kulkarni, R.V. and Venayagamoorthy, G.K., "An Estimation of Distribution Improved Particle Swarm Optimization Algorithm", Proc. IEEE/ ISSNIP,2007, pp. 539-544.
21. K.Indira, Dr. S. Kanmani , Gaurav Sethia.D, Kumaran.S, Prabhakar.J , 'Rule Acquisition in Data Mining Using a Self Adaptive Genetic Algorithm', In : First International conference on Computer Science and Information Technology, Communications in Computer and Information Science, 2011, Volume 204, Part 1, 171-178.
22. Ta-Cheng Chen, Tung-Chou Hsu, "GAs based approach for mining breast cancer pattern", Expert Systems with Applications, 2006: 674–681.
23. Zhan-min Wang, Hong-liang Wang, Du-wa Cui, "A growing evolutionary algorithm for data mining", 2nd IEEE international conference on information engineering and computer science, 2010:pp. 01-04
24. S. Dehuri, S. Patnaik, A. Ghosh, R. Mall, "Application of elitist multi-objective genetic algorithm for classification rule generation", Applied Soft Computing, 2008, pp. 477–487.

Authors Profile

K.Indira received the **B.E.** degree in Computer Science and engineering from Madurai Kmaraj University, Madurai, in 2004 and M.E in Computer Science and Engineering rom Annamali Univeristy in 2005 Currently doing her research in Pondicherry Engineering College, Puducherry. Her research interest includes data mining and Evolutionary algorithms

S.Kanmani received her B.E and M.E in Computer Science and Engineering from Bharathiyar University and Ph.D in Anna University, Chennai. She had been the faculty of Department of Computer Science and Engineering, Pondicherry Engineering College from 1992 onwards. Presently she is Professor in the Department of Information Technology, Pondicherry Engineering College. Her research interests are Software Engineering, Software testing, Object oriented system, and Data Mining. She is Member of Computer Society of India, ISTE and Institute of Engineers, India. She has published about 50 papers in various international conferences and journals.