Application of Enhanced Genetic Algorithm

To Symbolic Audubon Data

Hatem Elgothamy*
Department of Electrical & Computer Engineering
Oakland University
Rochester, Michigan, USA
*Email: hoelgoth [AT] oakland.edu

Mohamed A. Zohdy
Department of Electrical & Computer Engineering
Oakland University
Rochester, Michigan, USA

Hoda S. Abdel-Aty-Zohdy
Department of Electrical & Computer Engineering
Microelectronics & Bio-Inspired Systems Design Lab
Oakland University
Rochester, Michigan, USA

Hua Ming
Department of Computer Science
Oakland University
Rochester, Michigan, USA

Abstract—**Genetic algorithm (GA) is a rapidly growing area of Artificial Intelligence. This paper introduces a faster and less computationally expensive enhanced Genetic Algorithm (GA) than the standard GA. Five enhancements are introduced here (multiple weighted roulettes, multiple cross over points, multiple mates, utilizing the D4 wavelets and using normal distribution for selecting the initial population). Then the new enhanced GA was applied to a dynamic large optimization problem that uses symbolic data to differentiate between edible and poisonous mushrooms using twenty two different characteristics. Results was obtained using the enhanced GA as well as the standard GA. The enhanced GA was able to obtain the results using less number of calculations, which indicates it was less computationally expensive. A special java program was developed for this purpose; and Excel was used to represent the charts data.**

Keywords- genetic algorithm; bio inspired system; evolutionary algorithm; weighted roulette wheel; Daubechies wavelets; mushrooms; symbolic data; normal distribution;

I. INTRODUCTION

Genetic algorithm is an intelligent method for solving hard optimization problems in multiple dimensions. Genetic algorithm theory is based on Charles Darwin's theory of evolution that describes the principle "Survival of Fittest" for natural selection. Genetic algorithm imitates the process of evolution and follows the process of natural selection. Natural selection is probabilistic but favors the fittest individual in the generation. Enhancements are proposed to give a new variant of the standard genetic algorithm. The first step before running any genetic operations is the creation of the first generation. From that generation all the next generations will be derived one after the other. So the closer the first generation to a good solution, the less generation will need to be generated and the faster the system will be able to reach a solution with less computation.

Selection is the first genetic operation in the reproductive phase of GA. It helps the GA by directing the genetic search towards promising regions in the search space. The objective of selection is to choose the fitter individuals in the population that will create offspring for the next generation, commonly known as mating pool. The selected mating pool takes part in advancing the population to the next generation and hopefully close to the optimal solution. Selection pressure is a crucial factor that determines the efficiency of the algorithm and it is desirable that the mating pool should have good individuals to offer a better chance for a better offspring. The worthiness or the value of each individual depends on its fitness. The fitness value is determined by an objective function. The process of selection of individuals can be done using different algorithms. There is also improvement in the convergence velocity of the algorithm, which leads to reducing the time taken by the algorithm to reach the solution.

II. ENHAMCEMENTS

An important part of the selection process is the selection from one generation to another to create the basis of the next generation. The important requirement is that a set of the fittest individuals would have a higher chance of survival than a group of weaker ones. In nature, however, fitter individuals tend to have a better probability of survival and that will help them go on forward to form an updated mating pool for the next generation. Less fitting individuals should not be left without a chance though. Since in nature those individuals may have genetic codes that might still prove useful to future generations.

The first proposed enhancement is to use multiple weighted roulettes, as presented in [1]. This will help the selection pressure to be distributed from one generation to another. Depending on the application the roulettes can be used in parallel or in series.

Crossover has no fixed methods associated with it, but the basic general requirement is to carry the important genetic code from the parents forward to the next generation of offspring.

The second and third enhancements are to do the cross over process on multiple points as well as utilizing multiple mates to create offspring. Genetic material from all parents will be transferred to the offspring. The three possible scenarios resulting from this in regards to the fitness of the offspring, they can be fitter than their parents, has the same fitness or weaker than the parents. The highest fitness they have the higher their chances of survival are, and they will tend to die out if they have less fitness.

The fourth proposed enhancement is to utilize the Daubechies wavelets, which is named after its discoverer, the mathematician Ingrid Daubechies, as a preprocessing step. There are multiple types of Daubechies wavelets. An easy way to understand the Daubechies transforms is just to treat them as simple generalizations of the Daubechies D4 transform as presented in [2]. The length of the supports of the scaling signals is the most obvious difference between them. The most common Daubechies wavelets are shown in Fig.1.

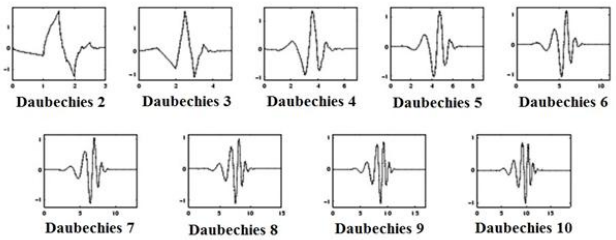

Figure 1. Daubechies wavelets

The D4 transform has four scaling function coefficients and can be extended to multiple levels as many times as the signal length can be divided by 2 as presented in [14]. The scaling function coefficients are:

$$\alpha_1 = \frac{1+\sqrt{3}}{4\sqrt{2}}, \quad \alpha_2 = \frac{3+\sqrt{3}}{4\sqrt{2}}, \quad \alpha_3 = \frac{3-\sqrt{3}}{4\sqrt{2}}, \quad \alpha_4 = \frac{1-\sqrt{3}}{4\sqrt{2}}. \tag{1}$$

All the scaling signals have the energy =1, which is an important property. This is because of the Euclidean norm of the α vector

$$\|\alpha\|2=1 \tag{2}$$

$$\alpha_1^2 + \alpha_2^2 + \alpha_3^2 + \alpha_4^2 = 1. \tag{3}$$

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = \sqrt{2}. \tag{4}$$

The most widely distribution used is the Gaussian normal distribution. Since the Gaussian distribution approximates many natural phenomena, it has developed into a standard of reference for many probability problems. That is why it was selected to be the fifth enhancement proposed to deal with the step of selecting the initial population.

Some of the characteristics of the normal distribution are that it is bell shaped, continuous and symmetric, for all values of X between -∞ and ∞ so that each conceivable interval of real numbers has a probability more than zero, as shown in Fig.2.

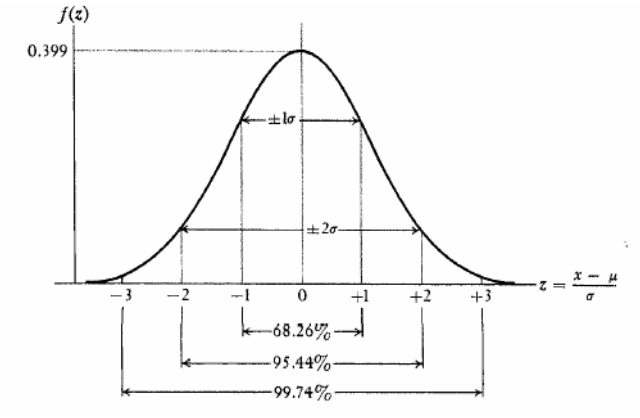

Figure 2. Normal (Gaussian) distribution

Gaussian distribution is actually a family of distributions since the shape of the distribution is determined by two parameters σ and μ. The normal density function rule is

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \tag{5}$$

About 95% of cases lie within 2 standard deviations of the mean, that is

$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = .9544 \tag{6}$$

Many things are actually normally distributed in real life, or very close to it. The normal distribution is also easy to work with mathematically. In a lot of practical cases, the methods developed using normal theory work very well even if the distribution was not normal. The larger the sample N the more the sampling distribution will approach the Gaussian form. Even if the population distribution was not normal, but if the sample N was large enough it can be approximated to the Gaussian distribution.

III. APPLICATION AND RESULTS

The mushroom database was constructed by the Audubon Society [9]. It includes descriptions of corresponding to twenty three hypothetical species samples of gilled mushrooms in the Agaricus and Lepiota Family. Each species is identified as poisonous or edible. Determining the edibility of a mushroom has no simple rule. The database identifies twenty two features to describe the smell, feel and looks, that can help determine if a mushroom is poisonous or edible.

The enhanced GA will be used to try to differentiate between edible and poisonous mushrooms using the group of 22 features. The list of characteristics used is shown in Table I. Each characteristic can have a group of different values that

range from two to twelve in some cases, and all these characteristics will add up to a total of 123 different attributes.

TABLE I. MUSHROOMS CHARACTERISTICS

1. Classes:
   - 1.1. Edible = e
   - 1.2. Poisonous = p
2. Cap-shape
   - 2.1. Bell = b
   - 2.2. Conical = c
   - 2.3. Convex = x
   - 2.4. Flat = f
   - 2.5. Knobbed = k
   - 2.6. Sunken = s
3. Cap-surface
   - 3.1. Fibrous = f
   - 3.2. Grooves = g
   - 3.3. Scaly = y
   - 3.4. Smooth = s
4. Cap-color
   - 4.1. Brown = n
   - 4.2. Buff = b
   - 4.3. Cinnamon = c
   - 4.4. Gray = g
   - 4.5. Green = r
   - 4.6. Pink = p
   - 4.7. Purple = u
   - 4.8. Red = e
   - 4.9. White = w
   - 4.10. Yellow = y
5. Bruises
   - 5.1. Bruises = t
   - 5.2. No = f
6. Odor
   - 6.1. Almond = a
   - 6.2. Anise = l
   - 6.3. Creosote = c
   - 6.4. Fishy = y
   - 6.5. Foul = f
   - 6.6. Musty = m
   - 6.7. None = n
   - 6.8. Pungent = p
   - 6.9. Spicy = s
7. Population
   - 7.1. Abundant = a
   - 7.2. Clustered = c
   - 7.3. Numerous = n
   - 7.4. Scattered = s
   - 7.5. Several = v
   - 7.6. Solitary = y
8. Gill-attachment
   - 8.1. Attached = a
   - 8.2. Descending = d
   - 8.3. Free = f
   - 8.4. Notched = n
9. Gill-size
   - 9.1. Broad = b
   - 9.2. Narrow = n
10. Gill-spacing
    - 10.1. Close = c
    - 10.2. Crowded = w
    - 10.3. Distant = d
11. Gill-color
    - 11.1. Black = k
    - 11.2. Brown = n
    - 11.3. Buff = b
    - 11.4. Chocolate = h
    - 11.5. Gray = g
    - 11.6. Green = r
    - 11.7. Orange = o
    - 11.8. Pink = p
    - 11.9. Purple = u
    - 11.10. Red = e
    - 11.11. White = w
    - 11.12. Yellow = y
12. Stalk-root
    - 12.1. Bulbous = b
    - 12.2. Club = c
    - 12.3. Cup = u
    - 12.4. Equal = e
    - 12.5. Rhizomorphs = z
    - 12.6. Rooted = r
    - 12.7. Missing = ?
13. Stalk-color-above-ring
    - 13.1. Brown = n
    - 13.2. Buff = b
    - 13.3. Cinnamon = c
    - 13.4. Gray = g
    - 13.5. Orange = o
    - 13.6. Pink = p
    - 13.7. Red = e
    - 13.8. White = w
    - 13.9. Yellow = y
14. Stalk-surface-below-ring
    - 14.1. Fibrous = f
    - 14.2. Scaly = y
    - 14.3. Silky = k
    - 14.4. Smooth = s
15. Stalk-color-below-ring
    - 15.1. Brown = n
    - 15.2. Buff = b
    - 15.3. Cinnamon = c
    - 15.4. Gray = g
    - 15.5. Orange = o
    - 15.6. Pink = p
    - 15.7. Red = e
    - 15.8. White = w
    - 15.9. Yellow = y
16. Stalk-surface-above-ring
    - 16.1. Fibrous = f
    - 16.2. Scaly = y
    - 16.3. Silky = k
    - 16.4. Smooth = s
17. Stalk-shape
    - 17.1. Enlarging = e
    - 17.2. Tapering = t
18. Veil-type
    - 18.1. Partial = p
    - 18.2. Universal = u
19. Veil-color
    - 19.1. Brown = n
    - 19.2. Orange = o
    - 19.3. White = w
    - 19.4. Yellow = y
    - 19.5. Two = t
20. Ring-number
    - 20.1. None = n
    - 20.2. One = o
21. Ring-type
    - 21.1. Cobwebby = c
    - 21.2. Evanescent = e
    - 21.3. Flaring = f
    - 21.4. Large = l
    - 21.5. None = n
    - 21.6. Pendant = p
    - 21.7. Sheathing = s
    - 21.8. Zone = z
22. Spore-print-color
    - 22.1. Black = k
    - 22.2. Brown = n
    - 22.3. Buff = b
    - 22.4. Chocolate = h
    - 22.5. Green = r
    - 22.6. Orange = o
    - 22.7. Purple = u
    - 22.8. White = w
    - 22.9. Yellow = y
23. Habitat
    - 23.1. Grasses = g
    - 23.2. Leaves = l
    - 23.3. Meadows = m
    - 23.4. Paths = p
    - 23.5. Urban = u
    - 23.6. Waste = w
    - 23.7. Woods = d

The database consists of 8,124 records with each record representing one sample. The samples are split into two classes, Edible (51.8% of the samples size or 4,208 records) and Poisonous (48.2% of the samples size or 3,916 records). There are a total of 2,480 missing attribute values, all of them for attribute #12 (Stalk-Root).

The program was developed using Java programming language. All obtained results in this paper are collected by running the program on the same computer without changing anything in its configuration or software.

The following Table II shows the results of trying to detect mushroom edibility using the two mentioned approaches. Each cell in the table below represents the average of 100 tests. For instance in Table II, the number of iterations when using the enhanced GA is 381.11. This number is obtained by calculating the average of running the same test 100 times. The first column in the table shows the number of iterations done when using the two approaches. The second column shows the average percentage of correct answers in those 100 tests performed while using the two approaches. The number of iterations needed in the standard GA to reach a result is almost double the enhanced GA, which means that the enhanced GA was able to outperform the standard GA and immensely reduce the amount of calculations without losing the quality of prediction.

TABLE II. NUMBER OF ITERATIONS

	Iterations	Correct
Enhanced GA	**381.11**	**97%**
Standard GA	**767.18**	**95%**

If any of the systems had to go through all iterations before reaching a decision, i.e. using brute-force. It would need to go through 186,852 iterations, and both systems were able to

reach a result using less than 800 iterations which saves a lot of processing and computing resources.

Fig. 3. Shows a comparison between the two rows of the first column in Table II. The percentage of correct solutions in Table II is also represented in Fig. 4. In conclusion, the proposed enhanced GA shows a significant improvement in computational cost compared to the standard GA with almost the same accuracy.

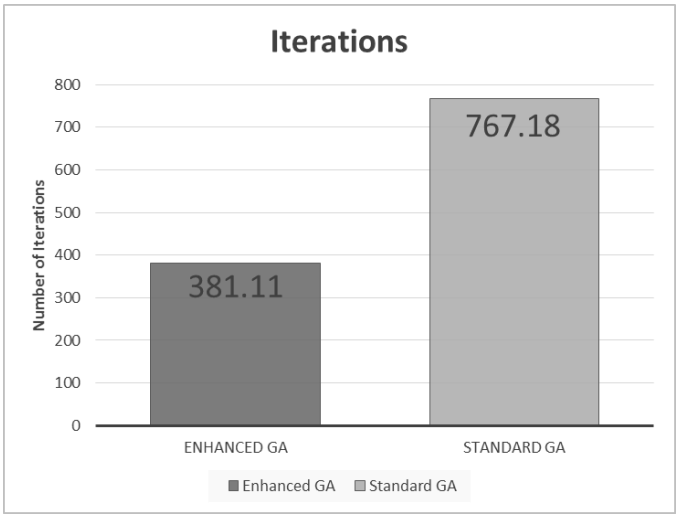

Figure 3. Comparing The Average Number Of Iterations needed to reach a result Using Standard & Enhanced Genetic Algorithms.

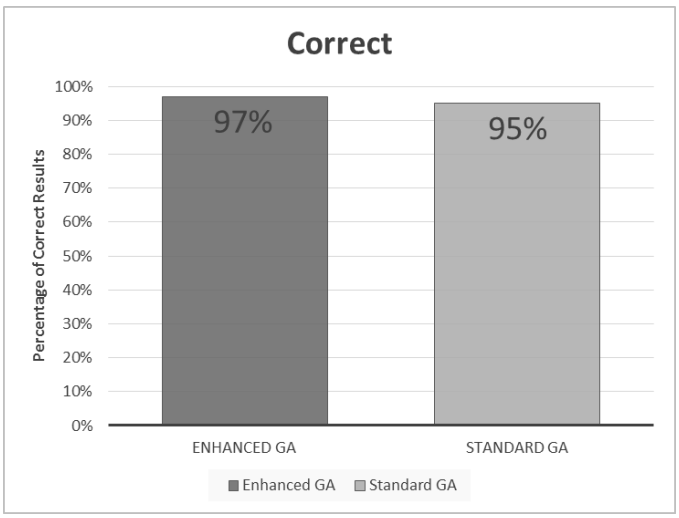

Figure 4. Comparing The Percentage of Correct Answers Using Standard & Enhanced Genetic Algorithms.

IV. CONCLUSION

This paper proposed a unique method for enhancing the computational cost of the system by introducing four enhancements to the standard GA which are (multiple weighted roulettes, multiple cross over points, multiple mates, utilizing the D4 wavelets and using normal distribution for selecting the initial population). Significant performance gains were observed utilizing the proposed methods when compared to the standard GA. To obtain significant performance improvement in computational cost over the standard GA, the proposed enhanced GA was developed. Through testing, the proposed enhanced GA is shown to be superior to the standard GA.

The number of iterations needed using the standard GA to reach a result, i.e. mushroom edibility, is almost double the enhanced GA which means that the enhanced algorithm was able to outperform the standard algorithm and immensely reduce the amount of calculations by almost half. Some background tasks are run by the operating system and the user usually has no control over them or the amount of memory and processing allocated to them, which might cause a small difference in the processing time

In some problems a less fitness can be accepted to save the computational cost, but that depends on the nature of the problem itself. In this paper, both genetic algorithm tests used a maximum number of 100 evolutions and a population size of 10. A higher number of allowed evolutions can lead to a higher fitness solution, but the question will be: does it worth the effort? Since enhancing the fitness by 0.1% can take double the time in some cases.

V. FUTURE WORK

As mentioned before, Genetic algorithm (GA) is a rapidly growing area of Artificial Intelligence. More enhancements to the GA are currently being researched, like islanding of special individual groups, using Chi Square to create the initial population, using annealing rate mutation with multiple sites and tournament for selecting the parents for the next generation.

The built Java program is always enhanced and updated to be a more comprehensive GUI and will be used with the enhanced genetic algorithm for more complex optimization problems.

REFERENCES

[1] H. Elgothamy, M. A. Zohdy and H. S. Abdel-Aty-Zohdy, "Design and application of an enhanced GA," 2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS), College Station, TX, 2014, pp. 864-867. doi: 10.1109/MWSCAS.2014.6908552

[2] Hatem Elgothamy,Mohamed A. Zohdy and Hoda S. Abdel-Aty-Zohdy, "Application of an Enhanced Genetic Algorithm to Radar System", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 7, Issue 1, January 2018. , pp. 37-43. DOI 10.17148/IJARCCE.2018.715

[3] S. Roth and H. S. Abdel-Aty-Zohdy, "Design and testing of D4 wavelets integrated chip preprocessor for chemical classifications," 2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS), College Station, TX, 2014, pp. 651-654. doi: 10.1109/MWSCAS.2014.6908499

[4] Garbacik, N.R., Zohdy, M.A., "Genetic Algorithm: Scattered Data Problems Using Lipschitz Interpolation", Aerospace and Electronics Conference, 2008, NAECON 2008. IEEE National, pp. 56-58.

[5] Ali Khan, A., Zohdy, M.A., "A Genetic Algorithm for Selection of Noisy Sensor Data in Multisensor Data Fusion", American Control Conference, 1997, pp. 2256-2262.

[6] A. Alfaro, M. Doan, J. Finke, M. Galdes, M. Zohdy, "Application of Divide and Conquer Extended Genetic Algorithm to Tertiary Protien Structue of CI-2", Jurnal of Applied Bionics and Mechanics 2007 Vol. 3 Issue 4.

[7] John H. Holmes, Ph.D., Dennis R. Durbin, M.D., M.S., Flaura K Winston, M.D., Ph.D., "Discovery of Predictive Models in an Injury Surveillance Database: An Application of Data Mining in Clinical Research", AMIA Symp. 2000, pp. 359-363.

[8] Robert E. Marmelstein, "Application of Genetic Algorithms to Data Mining", MAICS-97 Proceedings, 1997, pp. 53-57

[9] UCI Machine Learning Repository, audubon society mushrooms database, https://archive.ics.uci.edu/ml/datasets/mushroom retrieved January 1, 2018.

[10] Mithun K., "An Exordium of Genetic Algorithms", Mtech 2011.

[11] Holland, J.H., (1975) Adaptation in natural and artificial systems, University of Michican Press, Ann Arbor.

[12] Holland, J.H., (2000) Building blocks, cohort genetic algorithms, andhyperplane-dened functions, Evolutionary Computation, Vol.8 No.4, pp 373-391.

[13] John Henry Holland, "Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence", 1992.

[14] I. Daubechies, The wavelet transform, time-frequency localization and signal analysis IEEE Trans. Inf. Theory, 36 (5), pp. 961-1005, 1990.

[15] I. Daubechies, Wavelet transforms and orthonormal wavelet bases, pp. 1-33 in "Different perspectives on wavelets" AMS Short Course Lecture Note Series nr. 47, I. Daubechies (Ed.), AMS, 1993.

[16] A.R. Calderbank, I. Daubechies, W. Sweldens, and B.-L. Yeo, Wavelet transforms that map integers to integersAppl. Comp. Harm. Anal., 5 pp. 332-369, 1998.

[17] Chien-Feng Huang, A Study of Mate Selection Schemes in Genetic Algorithms—Part I, 200X by the Massachusetts Institute of Technology.

[18] Adam Lipowski and Dorota Lipowska, Roulette-wheel selection via stochastic acceptance, Physica A 391 (2012) pp. 2193-2196

[19] Robert E. Marmelstein, "Application of Genetic Algorithms to Data Mining", MAICS-97 Proceedings, Providence, Rhode Island, USA, 53-57

[20] Sastry, K., (2002) Evaluation-relaxation schemes for genetic and evolutionary algorithms. Master's thesis, University of Illinois at Urbana-Champaign, Urbana.

[21] Melanie Mitchel, Stephanie Forrest, John H. Holland, "The Royal Road for Genetic Algorithms", MIT Press 1991.

[22] Gilbert G. Walter, Xiaoping Shen. Wavelets and Other Orthogonal Systems, Second Edition , 2000 by CRC Press.

[23] Prem K. Kythe, Michael R. Schäferkotter. Handbook of Computational Methods for Integration , 2004 by Chapman and Hall

[24] James S. Walker. A Primer on Wavelets and Their Scientific Applications, 2002 by CRC Press

[25] Valliammal, N. Computer aided plant identification through leaf recognition using enhanced image processing and machine learning algorithms, Avinashilingam Deemed University For Women, 2013.

[26] Erik van Nimwegen, James P. Crutch¯eld, and Melanie Mitchell, "Statistical Dynamics of the Royal Road Genetic Algorithm", Theoretical Computer Science, 1998.

[27] Koza, J., (1992) Genetic Programming: on the programming of computers by means of natural selection, Volume 1, Complex adaptive systems, Bradford Books, 4th Edition, MIT Press, ISBN0262111705.

[28] Brian Carrigan, "Evolve Better Control Solutions With Genetic Algorithms", Microcontroller Central 2012.

[29] Goldberg,D.E., B.Korb, and K.Deb, (1989) Messy genetic algorithms: Motivation, analysis, and first results, Complex Systems, Vol. 3, pp 493-530.

[30] Jyotishree, Knowledge based operation and problems representation in genetic algorithms. Kurukshetra University, 2012.

[31] Adam Lipowski1 and Dorota Lipowska , "Roulette-wheel selection via stochastic acceptance", arXiv:1109.3627v1 [cs.NE] 16 Sep 2011.

[32] R.Sivaraj, Dr.T.Ravichandran, "An Efficient Grouping Genetic Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 21– No.7, May 2011.

[33] Prabha Verma and R.D.S. Yadava, "Genetic Algorithm Assisted Enhancement in Pattern Recognition Efficiency of Radial Basis Neural Network", Varanasi 221005, India, B.K. Panigrahi et al. (Eds.): SEMCCO 2011, Part I, LNCS 7076, pp. 257–264, 2011. © Springer-Verlag Berlin Heidelberg 2011.