Proposing PSO Based Algorithm for Classifying Breast Cancer Data Effectively

Pranjali Dewangan[1], Neelam Sahu[2]
[1,2]Dr. C.V.Raman University, Bilaspur (C.G.) India.

ABSTRACT
Digital Image Processing is processing of images that are digital in nature by a digital computer. Thresholding is the most commonly used intensity-based image segmentation technique which converts gray scale images into binary image. The performance of thresholding algorithms mainly depends on selection of threshold value. Various statistical properties such as maximum likelihood, moment, entropy and between class-variance has been utilized for selecting a proper threshold. This study takestwo objectives in account. The first to develop an efficient segmentation algorithm based on PSO and second to test proposed algorithm in classifying risk thereby classifying breast cancer data more effectively and efficiently. The proposed hybrid approach for data mining has included two phases. In the first phase, we adopted the statistical method in pre- processing. It can eliminate the insignificant features in order to reduce the complexity for next data mining stage. In the second phase, we proposed the data mining methodology that based on the standard PSO which called discrete PSO. In this study, we have used the Wisconsin breast cancer data set to test our proposed DPSO algorithm. In this study, a new hybrid approach of using both integrated statistical method and DPSO is proposed and successfully applied to the classification risk of Wisconsin- breast-cancer data set. According to our testing results, the proposed hybrid approach can improve the accuracy to 96.25%, sensitivity to 100% and specificity to 96.32%. These results are very promising compared to the previously reported classification techniques for mining breast cancer data.

Keywords:-Particle swarm Optimization, Optimization, Medical application

I INTRODUCTION

Digital Image Processing is processing of images that are digital in nature by a digital computer. Image Processing is motivated by three major applications; 1. improvement of pictorial info for human perception, 2. image processing for autonomous m/c application, 3. efficient storage and transmission. Segmentation is characterized as the process of dividing an image into distinct parts. Segmentation is utilized to identify an object or extract relevant information from digital images. Thresholding is a simple image segmentation methodology which yields a binary output from gray scale input images. Segmentation is the procedure by which an image is grouped into various units that are homogeneous with respect to one or more characteristics. It is an important task in image processing applications. The main objective of segmentation is to simplify the image into a form which is more meaningful and easier to analyse. Segmentation allows us to analyse an object or region in a more meaningful manner. Image segmentation can be done by three distinct methodologies. They are Intensity based segmentation, Edge-based segmentation, Region-based segmentation. Thresholding is the most commonly used intensity based image segmentation technique which converts gray scale images into binary image [Rutuparna Panda , Sanjay Agrawal, Sudipta huyan,2011].The performance of thresholding algorithms mainly depends on selection of threshold value. Various statistical properties such as maximum likelihood, moment, entropy and between class-variance has been utilized for selecting a proper threshold. It is defined as the partitioning of an image into non-overlapping regions that are homogeneous with respect to some visual feature, such as intensity or texture [Elsevier,2013]. This study taken into account following objectives.

(a) To develop an efficient segmentation algorithm based on PSO.
(b) To test proposed algorithm in classifying risk thereby classifying breast cancer data more effectively and efficiently.

II LITERATURE REVIEW

Segmentation algorithms are involved in virtually all computer vision systems, at least in a pre-processing stage, up to practical applications in which segmentation plays a most central role: they range from medical imaging to object detection, traffic control system and video surveillance, among many others. The importance of developing automated methods to accurately perform segmentation is obvious if one is aware about how tedious, time-consuming, subjective and error-prone manual segmentation can be according to the general principle on which the segmentation is based, we can build a taxonomy of the different segmentation algorithms distinguishing the following categories [M. Sonka, V. Hlavac, R. Boyle,R. Klette,2007]: thresholding techniques (based on pixel intensity), edge-based methods (boundary localization), region-based approaches (region detection), and deformable models (shape). Metaheuristic are general-purpose stochastic procedures designed to solve complex optimization problems[F. Glover,2003]. They are approximate and usually non-deterministic algorithms that guide a search process over the solution space. Unlike methods designed specifically for particular types of optimization tasks, they are general purpose algorithms and require no particular knowledge about the problem structure other than the objective function itself, when defined, or a sampling of it (training set) when the optimization process is based only on empirical observations. Metaheuristic are characterized by their robustness and ability to

exploit the information they accumulate about an initially unknown search space in order to bias the subsequent search towards useful subspaces. They provide an effective approach to manage large, complex and poorly understood search spaces where enumerative or heuristic search methods are inappropriate. Despite their importance and the number of scientific publications on the use of metaheuristic for deformable model optimization.

Liqiang Liu and Haijiao Ren et al(2010) in this paper, space contraction transformations are introduces into standard Ant Colony System algorithm to increase the speed and to improve the search ability of algorithm.

Myung-Eun Lee1 et al(2009) in this paper reviewer discuss about the ACO algorithm for the segmentation of brain MR images can effectively segments the fine details.

H. Shah-Hosseini (2011) in this paper, a novel metaheuristic called —Galaxy based Search Algorithm‖ or GbSA is introduced for multilevel thresholding. The proposed GbSA may be viewed as a variable neighbourhood search algorithm or as an Iterative local Search algorithm

This exhaustive Literature survey includes all relevant papers related with the hybridization of metaheuristic and segmentation models. At the same time, it aims at drawing some guidelines to help those who are willing to incorporate the advantages, and ease of use, of metaheuristic into the design of new segmentation approaches based on the same principles.

III RESEARCH METHODOLOGY

(a) The Proposed Hybrid Approach- The proposed hybrid approach for data mining has included two phases. In the first phase, we adopted the statistical method in pre- processing. It can eliminate the insignificant features in order to reduce the complexity for next data mining stage. In the second phase, we proposed the data mining methodology that based on the standard PSO which called discrete PSO. In this study, we have used the Wisconsin breast cancer data set to test our proposed DPSO algorithm. The data set included 9 features and 1 Order variable. We substituted the missing data by filling the values which appear the most frequently in that feature. Beside the Order variable, the value of 9 features is between 1 and 10, the higher value corresponding to a more unusual situation of the tumor such as the data in Table 1. The data set contains 699 points, 461 were diagnosed to be benign (Order = 2) and 238 to be metastatic (Order = 4). We divided the training data set which contains 459 patients' re- cords and validation data set which contains 240 patients' records from original data set randomly.

Table 1
The feature variable of dataset

Featurevariable	Domain	Simplifiedexpress
LumpViscosity	1–10	Z1
Cell Size Uniformities	1–10	Z2
Cell Shape Uniformities	1–10	Z3
Fringe Cohesion	1–10	Z4
SingleDeciduaCell Size	1–10	Z5
BasicCore	1–10	Z6
MildChromatin	1–10	Z7
RegularCore	1–10	Z8
Mitospore	1–10	Z9
Order	2,4	Z10
2:benign,4:metastatic		

IV DATA ANALYSIS

(a) Objectives - 1

- To develop an efficient segmentation algorithm based on PSO.

V DPSO FOR DATA MINING

Theparticleswarmoptimization(PSO)techniquei sapopulationbasedstochasticoptimizationtech niquefirstintroducedby Gordan, B., Armaghani, D. J., Hajihassani, M., & Monjezi, M. (2016). It belongs tothecategoryofSwarm

Intelligencemethods;itisalso anevolutionarycomputationmethodinspiredby themetaphorofsocialinteractionandcommunic ationsuchasbirdflockingandfish schooling.Thedetailshavebeen giveninthefollowing.

InPSO,asolutionisencodedas afinite-lengthstringcalleda particle(Allahverdi, A. (2015); Delice, Y., Aydoğan, E. K., Özcan, U., & İlkay, M. S. (2017); Gordan, B., Armaghani, D. J., Hajihassani, M., & Monjezi, M. (2016); Chen, K. H., Wang, K. J., Tsai, M. L., Wang, K. M., Adrian, A. M., Cheng, W. C., ... & Chang, K. S. (2014); Du, K. L., & Swamy, M. N. S. (2016); Shen, X., Chen, J. G., Zhu, X. C., Liu, P. Y., & Du, Z. H. (2015); Gonzalez-Vidal, A., Barnaghi, P., & Skarmeta, A. F. (2018); Shabbir, F., & Omenzetter, P. (2015)).Alloftheparticleshavefitnessvalues which are evaluatedby thefitnessfunctionto be optimized,andhave velocitieswhichdirectthe flyingof theparticles (Allahverdi, A. (2015); Delice, Y., Aydoğan, E. K., Özcan, U., & İlkay, M. S. (2017); Gordan, B., Armaghani, D. J., Hajihassani, M., & Monjezi, M. (2016); Du, K. L., & Swamy, M. N. S. (2016); Chen, K. H., Wang, K. J., Tsai, M. L., Wang, K. M., Adrian, A. M., Cheng, W. C., ... & Chang, K. S. (2014); Shen, X., Chen, J. G., Zhu, X. C., Liu, P. Y., & Du, Z. H. (2015); Dubey, A. K., Gupta, U., & Jain, S. (2015); Shabbir, F., & Omenzetter, P. (2015)).PSOis initializedwitha populationofrandomparticles,withrandomposi tionsandvelocities insidethe problemspace,and thensearchesfor optima byupdatinggenerations.Itcombineslocalsearcha ndglobalsearch yieldinginhighsearchefficiency.Each particlemovestowardsits bestpreviouspositionandtowardsthebestpartic leinthewhole swarmineveryiteration.Theformerisalocal bestanditsvalueis calledpbest, andthelatterisaglobalbestandits valueiscalled gbest inthe literature(Allahverdi, A. (2015); Delice, Y., Aydoğan, E. K., Özcan, U., & İlkay, M. S. (2017); Shen, X., Chen, J. G., Zhu, X. C., Liu, P. Y., & Du, Z. H. (2015); Gonzalez-Vidal, A., Barnaghi, P., & Skarmeta, A. F. (2018); Shabbir, F., & Omenzetter, P. (2015)).After findingthetwobestvalues,theparticle up-dates itsvelocityandpositionwiththefollowingequati onincontinuous PSO:

$$v^{t} \quad t-1 \quad t-1 \quad t-1 \quad t-1 \quad t-1$$

The values c1q1 and c2q2 determine the weights of two parts, and the value of (c1 þ c2) is usually limited to 4 (Gordan, B., Armaghani, D. J., Hajihassani, M., & Monjezi, M. (2016)).

Fig.6 The flowdiagramoftheproposedhybridapproach .

To apply PSO, several parameters including the number of population (m), cognition learning

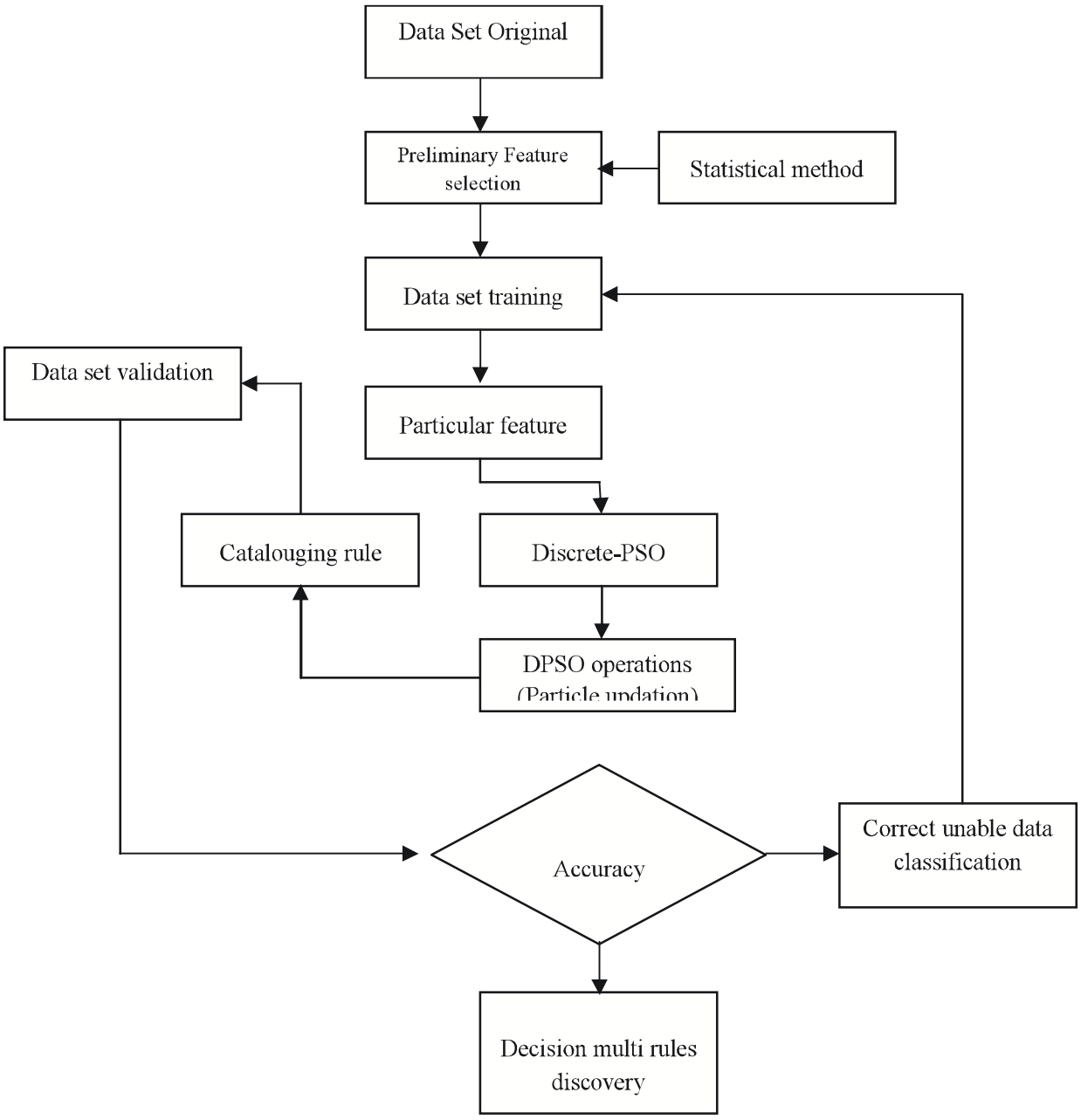
Data Set Original
Preliminary Feature selection
Statistical method
Data set training
Data set validation
Particular feature
Catalouging rule
Discrete-PSO
DPSO operations
Accuracy
Correct unable data classification
Decision multi rules discovery

Fig. 1

factor (c1), social learning factor (c2), inertia weight (w), and the number of iterations or CPU time should be determined properly. We conducted the preliminary experiments, and the complete computational procedure of the PSO algorithm can be summarized as follows:

(i) Initialize:Initializeparametersandpopulationwithrandom positionandvelocities.
(ii) Evaluation:Evaluatethefitnessvalue(the desiredobjective function)for eachparticle.
(iii) Findthe pbest:Ifthefitnessvalueofparticleiisbetterthan its bestfitnessvalue(pbest)in history,thenset currentfitnessvalueasthenewpbest to particlei.
(iv) Findthegbest:If any pbestisupdatedanditisbetterthanthe currentgbest, thenset gbestto thecurrentvaluc.
(v) Updatevelocity andposition:Updatevelocity andmovetothe nextpositionaccordingto Eqs.(1) and(2).
(vi) Stoppingcriterion:Ifthenumberofiteratio nsorCPUtimeare met,thenstop;otherwisegoback tostep2.

(a) Objectives - 2

To test proposed algorithm in classifying risk thereby classifying breast cancer data more effectively and efficiently.

Pre-processing using statistical method

From correlation and regression analysis, we could eliminate the insignificant features. In this phase, the feature ''Order" would be taken into a dependent variable and the remaining features would be taken into indepcndcnt variablcs. Tablc 2 shows thc experimental results from SPSS statistics package software. Obviously, we could find the three insignificant features including ''Fringe Cohesion", ''Single Decidua Cell Size" and ''Mitospore". In addition, the adjusted R2 has already reached 0.837 which represent the intersection between independent variablcs was insignificant. Hcncc, thc intcrscction could be ignored in this study.

Table 3 shows that we had the same results from correlation and regression analysis so we only need to kccp thcsc 6 fcaturcs to thc ncxt phasc which includcd ''Lump Viscosity", ''Cell Size Uniformities", ''Cell Shape Uniformities", ''Basic Core", ''Mild Chromatin and Regular Nucleoli.

Table 2
The experimental results from SPSS statistics package software

Modelsummary[b]

Model	R	R^2	AdjustedR^2	Std.
1	.925[a]	0.841	0.840	0.3844

Coefficients[b]

Model		Unstandardized		Standardized	t	Sig.	95%confidenceinterval	
		B	Std. error	Beta			Lowerbound	Upperb
1	(Constant)	1.509	0.038		45.942	0	1.449	1.577
	Z1	0.060	0.011	0.24	8.998	0	0.09	0.085
	Z2	0.048	0.018	0.149	3.499	0.007	0.07	0.16
	Z3	0.039	0.017	0.108	2.681	0.013	0.011	0.068
	Z4	0.020	0.013	0.040	1.421	0.163	−0.009	0.034
	Z5	0.017	0.08	0.039	1.406	0.169	−0.011	0.043
	Z6	0.097	0.014	0.358	14.495	0	0.15	0.115
	Z7	0.049	0.06	0.110	4.082	0	0.028	0.069

a Predictors: (constant), Z_9, Z_6, Z_1, Z_8, Z_5, Z_4, Z_7, Z_3, Z_2.
b Dependent variable: Z_{10}.

VI RESULT ANALYSIS AND CONCLUSION

The results of DPSO for mining the Wisconsin breast cancer data are tested. To perform the robustness of methodology in this study, we have presented the 10 results of experiment and the relative parameters of the algorithm below. The number of particle was 30, the number of generation was 50, Cw ¼ 0:1; Cp ¼ 0:4 and Cg ¼ 0:9. The setting of parameters in DPSO was case dependent, and we can do the research on them in the future study. In rule 1, we could derive the best accuracy to be 0.9528 that Cell Shape Uniformities > 2 and Mild

Chromatin > 1". In our study, the data could not be classified correctly by rule 1, we adopted the method of Nahar, J., Imam, T., Tickle, K. S., Ali, A. S., & Chen, Y. P. P. (2012) that new decision rule is to be explored. In this process, both the selected feature of training data not being classified correctly and all the unselected feature of data are preserved for mining in an additional rule (Nahar, J., Imam, T., Tickle, K. S., Ali, A. S., & Chen, Y. P. P. (2012)). After the repeated process, we found that Rule 2 is ''Lump Thick- ness > 3 and Basic Core > 2". So far, this study utilized two rules to improve the accuracy to 98.71%. Table 7 shows the comparison results with Gas. We found that the proposed DPSO can enhance the accuracy by 1.28%. Table shows the sensitivity can be up to 100% and specificity can be up to 98.21%, respectively. The performance of Type I error in GAs and DPSO to be equivalent. According to the above results, the proposed DPSO had shown to be better than the GAs in enhancing the performance of Type II error by 4.58%. Table 11 has compared the results of previous research in Wisconsin breast cancer with the proposed DPSO.The best way to improve the breast cancer victim's chance of long-term survival is to detect it as early as possible. Data mining and statistical analysis is one of the good solutions for searching the valuable information in large volumes of data (Muro, N., Larburu, N., Bouaud, J., Belloso, J., Cajaraville, G., Urruticoechea, A., & Séroussi, B. (2017, June)).

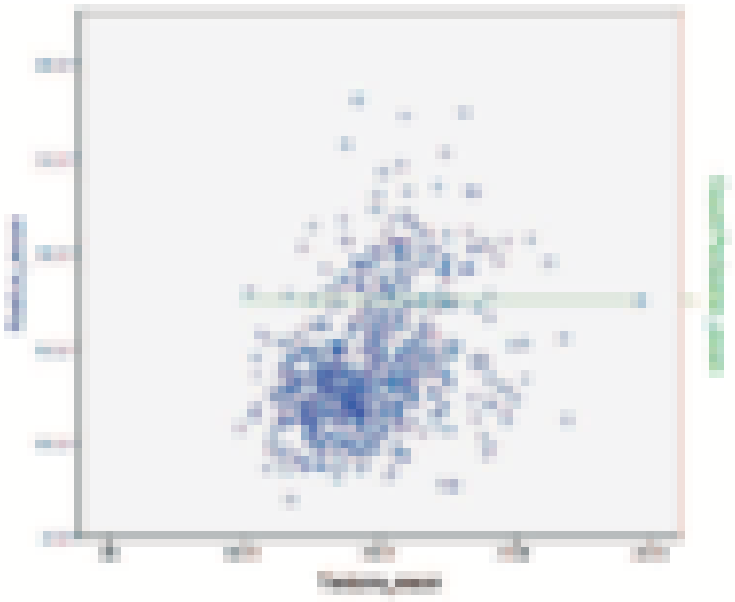

Fig. 2

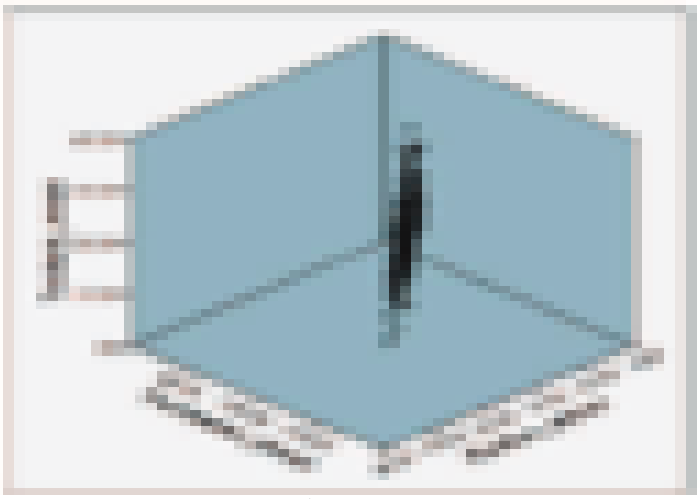

Fig. 3

In this study, a new hybrid approach of using both integrated statistical method and DPSO is proposed and successfully applied to the classification risk of Wisconsin- breast-cancer data set.

VII CONCLUSION

According to our testing results, the proposed hybrid approach can improve the accuracy to 96.25%, sensitivity to 100% and specificity to 96.32%. These results are very promising compared to the previously reported classification techniques for mining breast cancer data.

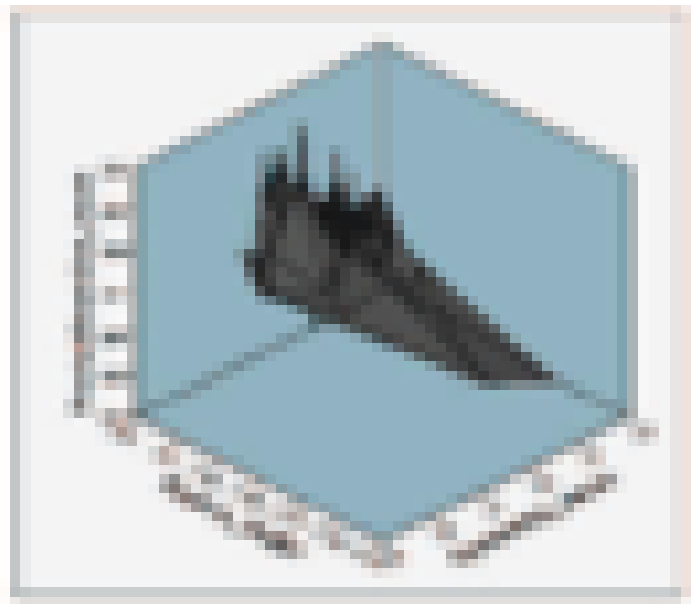

Fig. 4

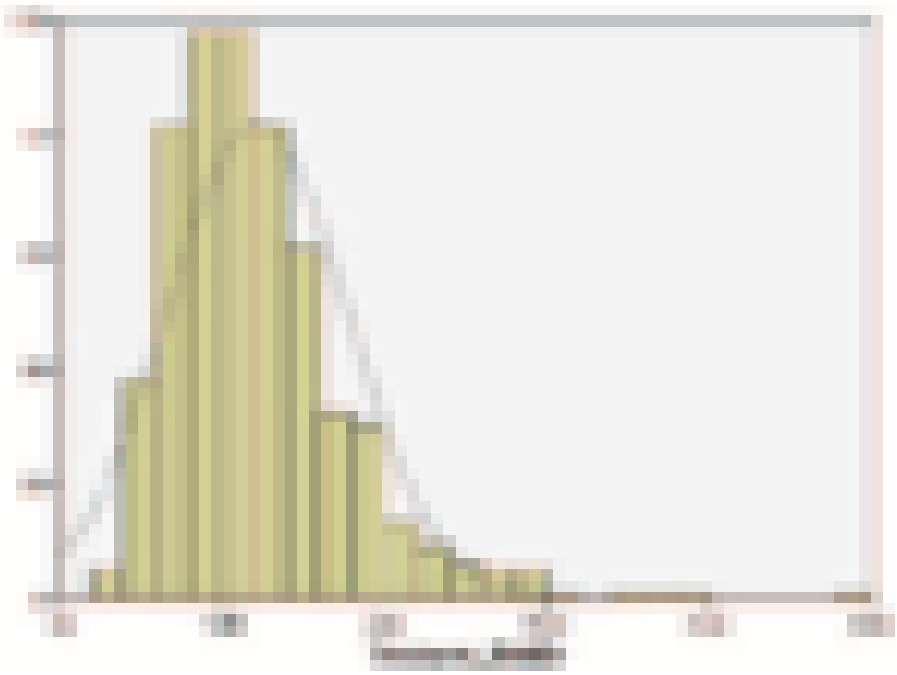

Fig. 5

Furthermore, the advantage of using statistical method to eliminate the insignificant features in pre-processing can improve the efficiency of DPSO process when the data set has included many features. Besides, the proposed DPSO can improve the constraint of genetic operation.

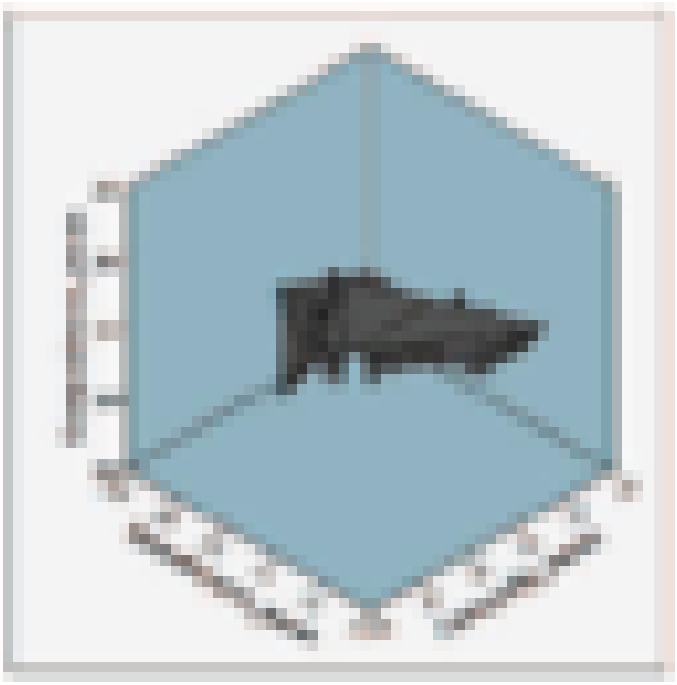

Fig. 6

The high classification accuracy from our proposed algorithm can be used as the reference for decision making in the hospital and the researchers. In future research, we recommend to better the process of data mining and apply it to the various domains in order to improve the medical quality for our lives.

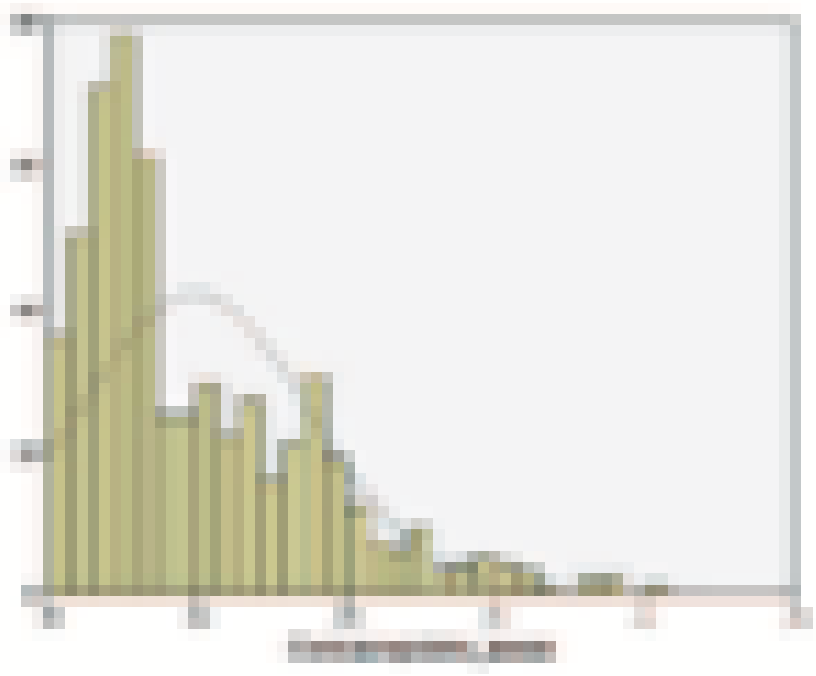

Fig. 7

REFERENCES

[1] Rutuparna Panda, Sanjay Agrawal, Sudipta Bhuyan, "Edge magnitude based multilevel thresholding using Cuckoo search technique.

[2] "Expert Systems with Applications 40 (2013) 7617–7628, 2013, Elsevier.

[3] M. Sonka, V. Hlavac, R. Boyle, Image Processing, Analysis, and Machine Vision, Thomson-Engineering,2007.

[4] R. Klette, Concise Computer Vision - An Introduction into Theory and Algorithms, Springer,2014.

[5] C. A. Floudas, P. M. Pardalos (Eds.), Encyclopedia of Optimization, Springer, 2009.

[6] D.-H. Li, M. Fukushima, On the Global Convergence of the BFGS Method for Nonconvex UnconstrainedOptimization Problems, SIAM Journal on Optimization 11 (4) (2000) 1054,1064.

[7] J. Grefenstette, Optimization of control parameters for genetic algorithms, IEEE Trans. Syst. Man Cybern.16 (1) (1986) 122–128.

[8] V. Zografos, Comparison of Optimisation Algorithms for Deformable Template Matching, in: Procs. Ofthe International Symposium on Advances in Visual Computing: Part II (ISVC),1097–1108, 2009.

[9] F. Glover, G. A. Kochenberger (Eds.), Handbook of metaheuristic, Kluwer Academic Publishers, 2003

[10] Sridevi, M, Mala, C. ; Sivasankar, E, "Optimized Multilevel Threshold Selection Using Evolutionary Computing,"17th International Conference on Network-Based Information Systems, IEEE, 2014.

[11] Paulo S. Rodrigues et al, "Improving a firefly meta-heuristic for multilevel image segmentation using Tsallis entropy", Pattern Anal Applications, 10044-015-0450-x, Springer, January 2015.

[12] Ye zhiwei et al, "Image Segmentation Using Thresholding and Artificial Fish-Swarm Algorithm", 2012 International

[13] Conference on Computer Science and Service System, 978-0-7695-4719-0/12, IEEE, 2012.

[14] V. Rajinikanth, N. Sri Madhava Raja, and K. Latha, "Optimal Multilevel Image Thresholding: An Analysis with PSO and BFO Algorithms", Aust. J. Basic & Appl. Sci., vol. 8, no. 9: pp. 443-454, 2014.

[15] Sathya, P.D. and Kayalvizhi, R. Optimal multilevel thresholding using bacterial foraging algorithm, Expert Systems with Applications, 38:15549 – 15564, 2011.

[16] Ming-Huwi Horng "Multilevel Thresholding selection based on the artificial bee colony algorithm for image segmentation "Expert Systems with applications, 2011.

[17] P.D. Sathya and R. Kayalvizhi, "Optimum Multilevel Image Thresholding Based on Tsallis Entropy Method with Bacterial Foraging Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010.