Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.1, 2019, pp. 69-73

A Proposed Method for Mining Breast Cancer Pattern Using Particle Swarm Optimization

Pranjali Dewangan[1] and Neelamsahu[2]
[1]Scholar, [2]Associate Professor
[1&2]Dr. C.V. Raman University, Bilaspur, Chhattisgarh, India
E-Mail: pranjalidewangan84@gmail.com

Abstract **- Breast cancer is one of the leading causes of death among women in many parts of the world. In this paper, we have developed an efficient hybrid data mining approach to separate from a population of patients who have and who do not have breast cancer. The proposed data mining approach has consisted of two phases. In first phase, the statistical method will be used to pre-process the data, which can eliminate the insignificant features. It can reduce the computational complexity and speed up the data mining process. In the second phase, we proposed a new data mining methodology, which based on the fundamental concept of the standard particle swarm optimization (PSO), namely discrete PSO. This phase aimed at creating a novel PSO in which each particle was coded in positive integer numbers and had a feasible system structure. Based on the obtained results, our proposed DPSO can improve the accuracy to 98.71%, sensitivity to 100%, and specificity to 98.21%. When compared with the previous research, the proposed hybrid approach shows the improvement in both accuracy and robustness. According to the high quality of our research results, the proposed DPSO data mining algorithm can be used as the reference for deciding on hospital and provide the reference for the researchers.**

Keywords: **Particle Swarm Optimization (PSO), Data Mining, Breast Cancer, Discrete PSO**

I. INTRODUCTION

Cancer is the cause of death on the planet. According to the data in the World Health Organization (WHO), we could find that over 10 million individuals have been clinically diagnosed to have cancer disorder and 6 million people will die to it in annually. As an instance, in the USA, cancer is the second prime death cause; it contributes to roughly 25 percent of mortalities. Prostate cancer in women frequently occurs, with more than 1 million new cases are diagnosed yearly. It is projected that approximately 500,000 women will die of breast cancer. In this paper, we present the pre-processing using the statistical method to remove the insignificant attributes in the feature selection procedure. The comparative research had demonstrated the absence of utilizing the statistical method in choice of feature subset factors that are statistically independent, so the learning algorithms might not be trained correctly [1]. The function of feature selection is especially essential when computational costly data mining tools are used, or if the information collection procedure is difficult or pricey [2]. In this paper, we have embraced the regression analysis. It might display the size of variance by the amount of square and correlation analysis which got in the size of inter-segmental between the attributes, i.e., "Class" along with another nine attributes. Considering that, the R2 of ANOVA was large enough to dismiss the intersection between all attributes. Thus, we could apply those as mentioned earlier, pre-processing statistical approaches to remove the insignificant attributes instead of select significant attributes. The purpose of the pre-processing measure is utilized to decrease the measurement of DPSO and improve the classification precision of breast cancer routine.

II. LITERATURE REVIEW

Breast cancer leads to death in many parts of the world among women. In 2007, roughly 178,480 girls in the USA were found to have invasive breast cancer. Approximately 40,970 girls will die from the disease annually [3]. On the other hand, the growth of medication technology makes mortality in breast cancer has diminished in the last ten years. Approximately 97 percent of girls can survive for five decades or longer. In this paper, we have studied to distinguish out of a population of individuals that have and who do not have breast cancer disorder [4], [5]. This analysis utilized the present data collection to establish the called principles, which can boost the validity of the analysis and help the treatment more efficiently.

Today, there has been an extreme Gain in the use of Data mining method in medication [6]. Because of the exponential expansion of the information technology and also the amount of information proliferates rapidly, we are very difficult to research the enormous amount of information [7], [8]. Data mining is just one of that algorithms commonly utilized to identify, validate, and forecast of information [9]. Lately, a data-mining tool that called "information classification method" has grown and implemented to the prediction of cancer disorder broadly [10]. The threat of information classification using the knowledge acquired from known historical data has been among the most intensively researched subjects in data, alternative science, performance research, and computer science [11], [12]. In previous studies, statistical strategies such as logistic regression and multi normal regression are the most commonly used data mining methods to assemble classification models. However, the intricacies of real-world classification issues are highly nonlinear. The experimental consequences of preceding studies demonstrated that these

methods are far superior compared to traditional approaches concerning forecast accuracy [3], [13], [14]. Hence, the Classification danger has received particular attention in utilizing soft-computing techniques. At the classification threat of breast cancer, the increase of mining Tool utilizing GA may have acceptable experimental outcomes. However, the new procedure that called people-based stochastic optimization (PSO) is superior compared to GA in many tasks, mostly in optimization regions [15]–[17]. This study enhances the method Festa, also Liu [18] that used a GAs for its classification procedure by restricting crossover and mutation. Since some receptor per chromosome is entirely equal, the difference of GAs is that the hunting space issue to be determined from the first solution of the algorithm. It means that the searching space is restricted, thereby reducing the potency of the algorithm.

III. THE PROPOSED HYBRID APPROACH

The hybrid approach for data mining has comprised of two phases. At the first stage, we embraced the statistical method of pre-processing. It may eliminate the insignificant attributes to reduce the complexity of following data mining stage. In the next phase, we suggested the data mining methodology, which depending on the typical PSO called discrete PSO. In this paper, the Wisconsin breast cancer data collection to check our DPSO algorithm has been utilized by us. The data set included nine attributes and one class factor. We substituted the missing data by filling the values, which appear the most often in that attribute. Besides the class factor, the value of 9 features is between 1 and 10, the higher value corresponding to some more unusual situation of this tumor, such as the data in Table I. The data set comprises 699 points, 458 were diagnosed to be benign (course = two) and 241 to be cancerous (class = 4). We divided the training data collection, which contains 466 patients' documents, and identification data set, which comprises patients' records from data placed randomly.

TABLE I DATA SET FEATURE VARIABLE

The feature variable of dataset		
Feature variable	**Domain**	**Simplified express**
Clump Thickness	1–10	X_1
Uniformity of Cell Size	1–10	X_2
Uniformity of Cell Shape	1–10	X_3
Marginal Adhesion	1–10	X_4
Single Epithelial Cell Size	1–10	X_5
Bare Nuclei	1–10	X_6
Bland Chromatin	1–10	X_7
Normal Nucleoli	1–10	X_8
Mitoses	1–10	X_9
Class	2,4	X_{10}
2:benign,4:malignant		

The instruction data set is used for learning breast cancer patterns and then generates the decision guideline (s). From Significance and regression analysis, we can remove the insignificant features. In this phase, the characteristic "Class" would be taken to a variable, and the features would be taken into Independent factors. Table II shows the experimental results in SPSS Statistics package program. Of course, we can find the three insignificant Features such as "Marginal Adhesion," "Single Epithelial Cell Size," and "Mitoses." Furthermore, 0.837 that has been reached by the adjusted R2 represent the junction between independent factors is insignificant. The intersection can be disregarded in this study.

TABLE II RESULTS FROM IBM SPSS SOFTWARE

Model summary[b]						
Model	**R**		**R^2 Std.**		**Adjusted R2**	**Error of the estimate**
1 Co-efficients[b]	.916[a]		0.839		0.837	0.3844
Model	**Unstandardized T Sig.**		**Standardized**		**95% confidence interval**	
	B	**Std. error**	**Beta**	**t**	**Lower Bound**	**Upper bound**
1(Constant)	1.506	0.033		45.935	0	1.442
X_1	0.064	0.007	0.19	8.993	0	0.05
X_2	0.045	0.013	0.143	3.493	0.001	0.02
X_3	0.033	0.012	0.104	2.672	0.008	0.009
X_4	0.011	0.008	0.034	1.412	0.159	-0.004
X_5	0.015	0.01	0.034	1.398	0.163	-0.006
X_6	0.093	0.006	0.353	14.485	0	0.08
X_7	0.041	0.01	0.106	4.075	0	0.021
X_8	0.035	0.007	0.113	4.719	0	0.021
X_9	0.005	0.01	0.009	0.506	0.613	-0.015

IV. THE PROPOSED DPSO

The suggested data mining methodology is based on the fundamental concept of the PSO, namely PSO. This paper aims at creating a novel PSO in which every particle has been coded from integer amounts and contains a system arrangement. So we only six attribute variables are staying after the evaluation stage, we have eliminated the insignificant feature factors. The debut of the suggested DPSO algorithm is provided in the next.

Encoding the idea of communicating was based on [19]–[21]; nonetheless, we have revised the original procedure to be systematized and effective in the method of solving the spell issues. The encoding in this analysis, we assume that the amount of chosen feature factors to be m. First, we specify the feasible solution that has a 1(3m + 1) array, which called an "individual" and the grids known as "cells."

The first cell represents the amount of feature factor; the second cell represents the factor where I = 1,..., m, and the third cell represents an inequality or equality, and the fourth cell reflects the threshold of the variable i. Besides the section of inequality or equality, we use the numeral to show what it means. Also, we define that the numeral"1" represents more significant about the numeral"2" signifies equal to and the numeral"3" represents smaller than.

The feeling of viable solution that are two attributes to be chosen, and the variable one at second position signifies that "Clump Thickness < Threshold 3", moreover, the factor 5 at fifth position signifies that "Single Epithelial Cell Size > Threshold two."

Fitness function in line with the situation of the upgrade, we could calculate. Concerning relative reference, we define that the TP, TN, FP, and FN since the rate parameters which is revealed. The amount of precision is the amount of the cancerous being chosen and also the sum of the benign the number of data will then divide them.

Besides the truth, we additionally consider the two performance measures: specificity and sensitivity. The definition of specificity, sensitivity, and precision are defined below.

$$Accuracy = \frac{TN + TP}{FN + FB + TN + TP}$$

$$Sensitivity = \frac{TP}{FN + TP}$$

$$Specifity = \frac{TN}{TN + FP}$$

The proposed individual update process is described below
1. Generate a random variable called n that is between 0 and 1.

2. If $0 \leq \xi < Cw$ is true then the original individual will be kept, else if $Cw \leq \xi < Cp$ is true then the original individual will be replaced by P best, else if $Cp \leq \xi < Cg$ is true then the original individual will be replaced by Gbest, else if $Cg \leq \xi \leq 1$ is true then the original individual will be replaced by new individual which generate randomly.

3. Repeat the above process until the terminative criteria has been met.

Updating the pace and rankings are the essential sections of PSO. They play a significant role in exchanging data. It contributes to a compelling mixture of partial solutions in different particles and accelerates the research process from the production stage.

From the conventional PSO, every particle should use over two equations, create three random numbers, five multiplications, and three summations to proceed to its next place. On the other hand, the suggested DPSO does not have to use the speed, it merely uses one arbitrary, two multiplications, and one contrast after Cw, Cp, and Cg are awarded. Hence, the suggested DPSO is significantly more efficient compared to other PSOs.

V. RESULTS AND DISCUSSION

Table III shows the results of DPSO for mining Wisconsin breast cancer information. To execute the robustness of methodology in this study, we have introduced the ten results of the experiment and the relative parameters of this algorithm below.

The amount of particle was 30, the number of production was 50, Cw 1/4 0:1; Cp 1/4 0:4 and Cg 1/4 0:9. The setting of parameters in DPSO was case reliant, and we can research these in future research.

In our study, the data could not be classified correctly by rule 1, we adopted the method of Chen and Hsu (2006) that the new choice rule is to be researched. Within this process, the chosen feature of training data not being categorized correctly and all the unselected characteristic of data is maintained for mining in a new rule (Chen & Hsu, 2006).

Following the repeated procedure, we discovered that Rule 2 is" Clump Thickness > 3 and Bare Nuclei > 2". Thus far, this study utilized two rules to enhance the precision to 98.71%.

According to the above effects, the suggested DPSO had shown to be better compared to GAs in enhancing the operation of Type II error by 4.58 percent. Table IV has contrasted the results of research in Wisconsin breast cancer using the DPSO.

TABLE III EXPERIMENTAL RESULTS

Item	Rules	Training accuracy	Validation accuracy	No. of particle	No. of generation	Time(s)
Rule 1						
1	X_2>2andX_3>1	0.9442	0.9227	30	50	24.104
2	X_2>2andX_5>1	0.9356	0.9013	30	50	23.634
3	X_2>3andX_3>2andX_7>1	0.912	0.9528	30	50	23.774
4	X_3>2andX_7>2	0.9356	0.9528	30	50	23.483
5	X_3>1andX_6>1	0.9378	0.9356	30	50	23.314
6	X_2>1andX_3>1andX_6>1	0.9421	0.9356	30	50	23.253
7	X_2>1andX_3>2andX_7<10	0.9206	0.9056	30	50	24.155
8	X_3>2andX_7>2	0.9356	0.9528	30	50	23.433
9	X_2>2andX_3>1	0.9442	0.9227	30	50	23.845
10	X_3>1andX_6>1	0.9378	0.9356	30	50	24.535
	Avg.	0.93455	0.93175	30	50	23.753
	Std.	0.010385	0.018704	0	0	0.41331
Rule 2						
1	X_1>3andX_6>2	0.9505	0.9828	10	30	3.475
2	X_1>3andX_6>2	0.9505	0.9871	10	30	3.555
3	X_1>3andX_6>2	0.9505	0.9871	10	30	3.575
4	X_2>1andX_6>1	0.9474	0.9785	10	30	3.615
5	X_3>1andX_6>2	0.9567	0.9785	10	30	3.736
6	X_1>3andX_6>2	0.9505	0.9871	10	30	3.676
7	X_6>2andX_8>1	0.9412	0.9785	10	30	3.485
8	X_1>3andX_6>2	0.9505	0.9871	10	30	3.605
9	X_6>2andX_7>3	0.9443	0.9742	10	30	3.515
10	X_1>3andX_6>2	0.9505	0.9871	10	30	3.435
	Avg.	0.94926	0.9828	10	30	3.5672
	Std.	0.004185	0.004965	0	0	0.09411

TABLE IV COMPARISON BETWEEN PREVIOUS RESULTS AND DPSO

	DPSO (this study proposed)	Gas (Chen & Hsu, 2006)	DPSO (Sousa *et al.*, 2004)	Constricted-PSO (Sousa *et al.*, 2004)	Linear decreasing Weight-PSO (Sousa *et al.*, 2004)	J48
Accuracy	98.71%	96.995%	94%	93.4%	93%	92.9%

VI. CONCLUSION

The very best method to improve the breast cancer sufferer's likelihood of long-term survival would be to detect it as soon as possible. In this paper, a new hybrid method of using both incorporated statistical method and DPSO has been suggested and successfully applied to the classification risk of Wisconsin breast- cancer data collection. By our testing results, the proposed hybrid strategy may enhance the accuracy to 98.71%, sensitivity to 100%, and specificity to 98.21 percent. These results are auspicious when compared with the previously mentioned classification methods for mining breast cancer information. Furthermore, the advantage of using a statistical process to eliminate the insignificant features in pre-processing can improve the efficiency of DPSO process once the data set has included many features. Besides, the restriction of operation that is genetic can improve in GAs. The high classification accuracy from our proposed algorithm can be utilized as the reference for decision making in the researchers along with the hospital. In a future study, we not only continued to ameliorate the process of data mining but applied it to the many domains in order to improve the health quality for our own lives.

REFERENCES

[1] N. Padhy and R. Panigrahi, "An efficient approach of Multi-Relational data mining and statistical technique," *in Advances in Intelligent Systems and Computing*, Vol. 327, pp. 99–111, 2014.

[2] H. De Weerd, R. Verbrugge, and B. Verheij, "Agent-based models

for higher-order theory of mind," *in Advances in Intelligent Systems and Computing*, Vol. 229 AISC, pp. 213–224, 2014.

[3] A. K. Dubey, U. Gupta, and S. Jain, "A survey on breast cancer scenario and prediction strategy," *in Advances in Intelligent Systems and Computing*, Vol. 327, pp. 367–375, 2014.

[4] K. Chen, F. Y. Zhou, and X. F. Yuan, "Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection," *Expert Syst. Appl.*, 2019.

[5] Y. Liu and Y. Y. Chung, "Mining cancer data with discrete particle swarm optimization and rule pruning," *in ITME 2011 - Proceedings: 2011 IEEE International Symposium on IT in Medicine and Education*, 2011.

[6] J. D. P. Rao and R. K. Akuli, "A Brief Study on Measures to Improve Cyber Network Security," pp. 20–22, 2015.

[7] J. D. P. Rao and A. Srivastava, "Impact of Web Enabled Knowledge Platform: An Analysis," *Int. J. Comput. Sci. Manag. Syst.*, Vol. 4, No. 1, pp. 1–7, 2012.

[8] R. K. Akuli, J. D. P. Rao, and S. Kurariya, "A Study Of Security Mechanisms Implemented In Network Protocols," *Indian Streams Res. J.*, Vol. 5, No. 11, pp. 1–3, 2015.

[9] Y. Zhang, S. Wang, and G. Ji, "A Comprehensive Survey on Particle Swarm Optimization Algorithm and Its Applications," *Mathematical Problems in Engineering*, 2015.

[10] I. C. Trelea, "The particle swarm optimization algorithm: Convergence analysis and parameter selection," *Inf. Process. Lett.*, Vol. 85, No. 6, pp. 317–325, Mar. 2003.

[11] A. Unler and A. Murat, "A discrete particle swarm optimization method for feature selection in binary classification problems," *Eur. J. Oper. Res.*, 2010.

[12] L. F. Chen, C. T. Su, and K. H. Chen, "An improved particle swarm optimization for feature selection," *Intell. Data Anal.*, 2012.

[13] H. K. Feng, J. S. Bao, and Y. Jin, "Particle swarm optimization combined with ant colony optimization for the multiple travelling salesman problem," *in Materials Science Forum*, Vol. 626- 627, pp. 717–722, 2009.

[14] Z. H. Zhan and J. Zhang, "Discrete particle swarm optimization for multiple destination routing problems," *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 5484 LNCS, pp. 117–122, 2009.

[15] P. Ghamisi, M. S. Couceiro, N. M. F. Ferreira, and L. Kumar, "Use of Darwinian Particle Swarm Optimization technique for the segmentation of Remote Sensing images," *in International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 4295–4298, 2012.

[16] A. E. Hassanien, H. M. Moftah, A. T. Azar, and M. Shoman, "MRI breast cancer diagnosis hybrid approach using adaptive ant-based segmentation and multilayer perceptron neural networks classifier," *Appl. Soft Comput. J.*, Vol. 14, No. PART A, pp. 62–71, 2014.

[17] S. Liu, X. Wang, and X. You, "Cultured differential particle swarm optimization for numerical optimization problems," *in Proceedings - Third International Conference on Natural Computation, ICNC 2007*, Vol. 4, pp. 642–646, 2007.

[18] N. Liu and W. Song, "A PCA-based algorithm for Kalman filtering in colored noise environments," *Gaojishu Tongxin/Chinese High Technol. Lett.*, Vol. 24, No. 5, pp. 520–524, 2014.

[19] M. Kumari and V. Singh, "Breast Cancer Prediction system," *in Procedia Computer Science*, Vol. 132, pp. 371–376, 2018.

[20] T. Sousa, A. Silva, and A. Neves, "Particle Swarm based Data Mining Algorithms for classification tasks," *Parallel Comput.*, Vol. 30, No. 5–6, pp. 767–783, May 2004.

[21] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artif. Intell. Med.*, Vol. 34, No. 2, pp. 113–127, Jun. 2005.