Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 9, Issue. 5, May 2020, pg.113 – 118

Modified Genetic Folding Algorithm for Breast Cancer Classification Dataset

Mohammad A. Mezher
Dept. of Computer Science, Fahad Bin Sultan University, KSA, mmezher@fbsu.edu.sa

Abstract— Cancer is a disease that develops in the human body due to gene mutation. Because of various factors, cells can become cancerous and grow rapidly, destroying normal cells at the same time. Support vector machines allow for accurate classification and detection of the classes. The advantage of kernel selection is to derive global learning rates for SVMs using the Genetic Folding algorithm. The developed GF algorithm outperforms traditional SVMs in the UCI Breast Cancer Wisconsin Diagnostic (BCWD) dataset under a certain comparative analysis, which is conducted under a set of conditions that describe the behavior of the compared algorithms. The observation that relates the GF performance appears to be comparable with SVM. The statistical analysis relies on a careful analysis of the ROC curve. Moreover, the GF algorithm shows that accuracy rates are obtained adaptively, that is, without knowing the parameters resulting from the margin conditions. The experimental results show that the one GF operator produces superior classification accuracy. The proposed method plays an important role in the detection of breast cancer in an efficient time frame.
Index Terms— GF, Breast Cancer Wisconsin, UCI, classification, SVM, Evolutionary Algorithms, Genetic Folding.

I. INTRODUCTION

In recent years, breast cancer has become one of the growing causes of death in women, and the survival rate is low. Researchers are at an early stage in exploring different kinds of approaches for breast cancer detection. The presence of this disease is identified by multiple diagnostic tests and procedures; a biopsy taken from the breast is one of the most common. Increasing numbers of cancer cells are caused by various factors, including noise, food quality, and smoking. In the Middle East, the rate has been rising for more than two decades [2]. In 2016, there were 45,980 new cases and 20,063 deaths in the region [1]. Therefore, developing new models and improving existing models has become an urgent need.

The Breast Cancer Wisconsin Diagnostic (BCWD) dataset can provide very useful information for research in the medical area, based on previous data and experience. Among various cancer cell detection and classification methods, machine-learning algorithms are useful for information retrieval. ML collects useful information from a huge data set based on the premise of past encounters and defines complicated architecture.

The Support Vector Machine (SVM) was introduced by Cortes and Vapnik as a learning machine based on two-group classification problems. The statistical learning theory solves problems of classification and regression by mapping the nonlinear input vectors to a very high-dimensional feature space [3]. Recently, SVM applications have been booming in popularity in the field of biomedical research, due to the effectiveness of their predictive abilities and their ability to classify data. SVM has been employed in various classification problems and is currently of interest in breast cancer detection due to its robustness. However, selecting these kernel functions has some limitations, such as being unpredictable, time consuming, and not suitable for a specific dataset. In SVM, a variety of kernel functions result in dissimilar accuracy. However, only a small amount of literature has focused on examining the predictive performances of SVM based on the kernel-evolutionary algorithm—specifically, using the GF algorithm.

However, the SVM is a kernel-based approach, whereby kernel selections are the core of the machine. The regularization parameter and kernel function selections are the primary tasks to determine in SVM before conducting training. The goal of

SVM is to increase precision with a maximum margin, as compared to the best predetermined SVM kernels. Currently, researchers engaged in using the SVM for breast cancer detection utilized SVM approaches such as SVM for ICS [4], GABC [5], or twelve different SVMs [6].

For this research, in terms of testing the accuracy of classification, a new approach has been proposed based on SVM and the Genetic Folding algorithm. Here, a comparison of newly generated SVM kernels, using the Genetic Folding algorithm with different genetic operators, was conducted for the WBCD dataset. The Genetic Folding (GF) algorithm is useful for improving prediction of the disease. GF, which is mainly based on the evolutionary concept, used to generate the best kernel model for classifying and predicting classes. Previously, different GF research was done using various analytic approaches, such as ROC curve and structural complexity. Related works are highlighted in Section II. In Section III, a background about the GF toolbox is discussed. In Section IV, the pre-processing, proposed method and experimental results and analysis are discussed. Section V consists of the conclusion and future scope of the proposed work.

II. LITERATURE REVIEW

The accuracy of classification tests largely depends on the expertise of the radiologist; consequently, there is a chance that a malignant lesion is diagnosed as benign or vice versa. There are a good number of studies conducted with many computational intelligence techniques via GA-SVM [7]. The results obtained in the paper shows GA-SVM was very close to Random Forest (RF), with the back propagation neural network model and Extreme learning machine coming in behind it.

A model in [8] includes different Parallel Genetic Algorithm (PGA) versions to build a classifier of detecting breast cancer in the UCI dataset. The study simultaneously investigates the use of PGA and a Coarse-Grained Parallel Genetic Algorithm (CGPGA) in improving SVM classification performance. Whereas in [9], the paper addressed the problem of feature selection using a GA-SVM classifier on the WBCD dataset. The research shows that the use of a best fitness function of an SVM classifier improves the evaluation performances of GA. Whereas in [10], GA was used to select the most separable features before training the BCW dataset using SVM.

In [11], they used SVM and SVM ensembles for small and large-scale breast cancer datasets to compare the prediction accuracy. In a small-scale dataset, the results show that SVM ensembles using both linear and Radial Basis Function (RBF) kernels are the better choices. For a large-scale dataset, SVM ensembles using RBF kernels outperform other classifiers.

The kernel selections remain a challenge for SVM optimization problems. The kernels used in [12] include linear, polynomial, and RBF and Genetic Programming (GP). Based on the "10-fold cross validation process", they compared the results of a linear GP approach to SVM and concluded that GP is more accurate than other machine learning algorithms.

In [13], a combination method of RF and SVM classifier is proposed for early diagnosis of WBCD. The paper performed different experiments for train-test data ratio and achieved results of more than 98% accuracy.

III. GENETIC FOLDING TOOLBOX

The Genetic Folding [14] toolbox was first introduced in [15]. GF starts by generating the scalar root, which must be an operator with a style of two operands. The descendant cells of the GF root will be represented by linear floating numbers. On every GF cell, a floating number will be produced either with two operands – left and right numbers along with a period or with one operand – or a right number only, along with a period. The two operands consist of left child and right child, whereas the one operand has only right child.

In each generation, GF focuses on the fittest GF individual attribute of the operators and operands. This is the representation of the SVM kernel of the classifier. In a number of generations, GF will be able to produce a fittest kernel with the smallest error value or highest accuracy rate.

IV. PROPOSED MODEL

This evolutionary algorithm produces new individuals from the concept of integrating each operator in the subset with corresponding operands. GF starts making a pool of individuals set (S). In every generation, GF generates randomly, by the selected attributes, valid GF individuals. Also, even when considering only a small subset of operators, GF was able to generate reliable kernel models [17]. GF starts producing an initial population from a subset (S). GF:

- Starts by defining specific parameters of GF length, mutation rate, crossover rate, maximum generations, population size, cross validation value, and operator list: see Table I
- Invokes an initial population function to generate a specific number of GF individuals (kernels)
- Scales the WBCD dataset to be in the same range
- Creates an SVM classifier using produced custom GF kernel
- Calculates fitness of produced custom GF kernels
- Maintains the best kernels found in the initial populations

- Plots values found in the generation using fitness, structure complexity, population diversity, and accuracy versus complexity figures

Working steps of GF producing generations uses these following evolutionary components:

1. Invokes a selection method where an original population remains as long as it is fitter than the new one, otherwise some replacements will take place if a generated random number was less than their fitness value
2. Applies crossover operator only in half of the population
3. Applies mutation operator only in the other half of the population
4. Calculates fitness values of each half and saves the fittest kernel individuals
5. Updates all figures with the new produced values
6. Repeats Steps 1 to 5 until a predefined number of generations is reached

Working steps of GF algorithm producing kernels (individuals) are guided by the following rules:

- Every gene in the (S) subset belongs to the float number, which represents the operators (+, -, *), or else the gene is turned into a leaf and labelled with an integer number
- The first gene (root) of GF individuals generates using only the scalar operators (Plus_s, Minus_s and Multi_s)
- GF then generates descendants (leafs except root) using scalar and vector operators (Plus_v, and Minus_v)
- The maximum numbers of operators produced in GF individuals are no more than the following equation:

$$Oplimit = int((GF\_length - 3) / 4) \quad (1)$$

V. RESULTS AND ANALYSIS

In this paper the UCI Breast Cancer Wisconsin Diagnostic dataset is conducted. The BCWD dataset contains features projected from a digitized image of a fine needle aspirate (FNA) of a breast figure. The attributes columns in the dataset describe characteristics of the cell nuclei presented in the image.

There are 37% malignant breast tumor cases and 63% benign tumor cases being studied. The experiment compiled data using Python JetBrains PyCharm CE and ran it with a CPU with 1.80 GHz and 1.99 GHz Intel Core i7-8550U CPU and 16 GB RAM.

The SVM-GF algorithm was applied over the datasets. Table I shows parametric values selected for the experiment. The kernel selection was performed using the generation schema specified in Section III. The SVM-GF results provide an accuracy of between 96% and 98.8% on test set with 5 cross-validation. There was an increase in accuracy obtained with kernels having a length of more than 17 genes and 8 folds. Weak behavior of GF was found where genetic operators' rates were very small. RBF, linear, and polynomial kernels outperformed GF where rates were small. In this case, the accuracy obtained for both a small and large number of genes is 96%. The data for both training and testing was selected randomly from the complete set for all the experiments.

TABLE I. PARAMETRIC VALUES

Parameter type	***Exper. 1***	***Exper. 2***	***Exper. 3***	***Exper. 4***
No. of generations	20	20	20	20
No. of population	50	100	50	50
Xover rate	0.5	0.5	-	0.9
Mutation rate	0.01	0.01	0.7	-
Cross validation	5	5	5	5
GF length	20	30	20	20
Operators	'Minus_s','Plus_s', 'Minus_v', 'Plus_v', 'Multi_s',			
Operands	'x', 'y'			

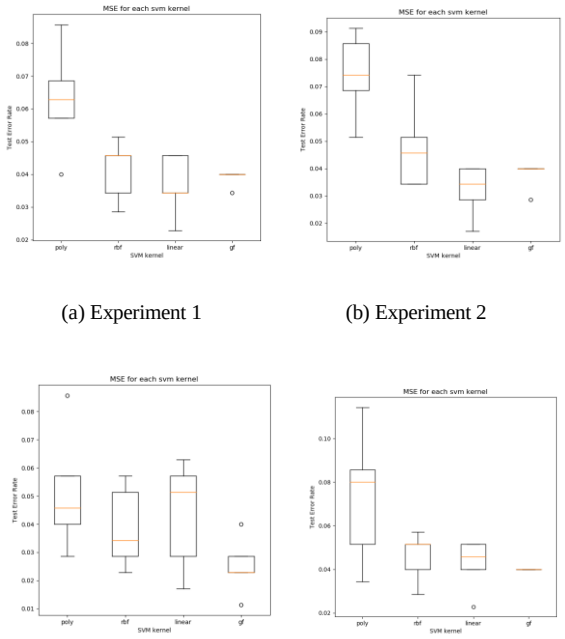
- means no GF operators were used

Using the above optimized parameters to train the model with our dataset, the results obtained are found in Table II. The obtained results show for all the cases with numbers of mutation rates 0.01, 0.7, or without a mutation operator. Also, Table II demonstrates a promising trend in the results, with an average accuracy of 98.8%.

TABLE II. GF ACCURACY AND AREA UNDER CURVE RESULTS

Experiment no.	***Accuracy***	***AUC***
Experiment 1	96.0	0.99
Experiment 2	96.1	0.99
Experiment 3	**98.85**	0.99
Experiment 4	96.0	0.99

Figure 1 shows the Mean Square Error (MSE) of the conducted experiments. At the red dot, the sensibility and specificity are simultaneously higher. The AUC is calculated as 0.99%, which proves that the GF-SVM is effective and has a good result in all sorts of operators.



(a) Experiment 1 (b) Experiment 2

(c) Experiment 3 (d) Experiment 4

Fig. 1. Mean Square Error of four experiments

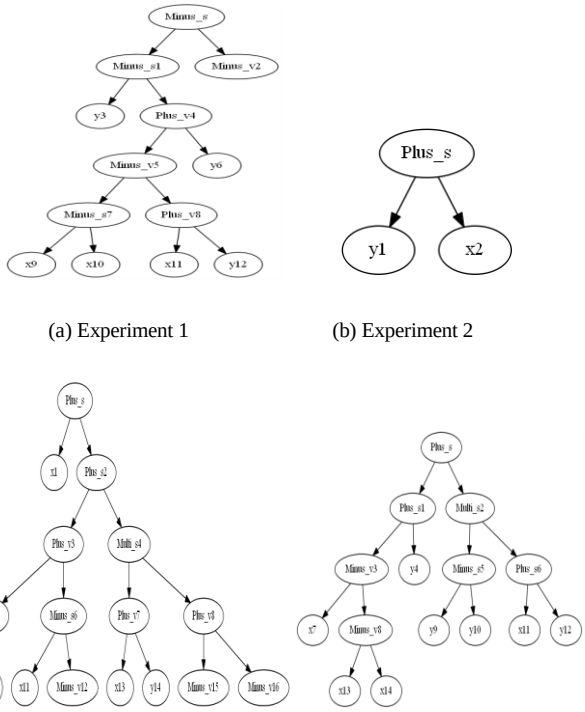
Table III contains the results of both GF strings and indices of the best individuals found. In Experiment 2, the GF string was very small, but just as efficient as the GF individual found in Experiment 1.

TABLE III. GF BEST INDIVIDUALS

Experiment 1	GF String	['Minus_s', 'Minus_s', 'Minus_v', 'y', 'Plus_v', 'Minus_v', 'y', 'Minus_s', 'Plus_v', 'x', 'x', 'x', 'y']
	GF Index	['1.2', '3.4', '0.2', '0.3', '5.6', '7.8', '0.6', '9.10', '11.12', '0.9', '0.10', '0.11', '0.12']
Experiment 2	GF String	['Plus_s', 'y', 'x']
	GF Index	['1.2', '0.1', '0.2']
Experiment 3	GF String	['Plus_s', 'x', 'Plus_s', 'Plus_v', 'Multi_s', 'Plus_s', 'Minus_s', 'Plus_v', 'Plus_v', 'x', 'x', 'x', 'Minus_v', 'x', 'y', 'Minus_v', 'Minus_v']
	GF Index	['1.2', '0.1', '3.4', '5.6', '7.8', '9.10', '11.12', '13.14', '15.16', '0.9', '0.10', '0.11', '0.12', '0.13', '0.14', '0.15', '0.16']
Experiment 4	GF String	['Plus_s', 'Plus_s', 'Multi_s', 'Minus_v', 'y', 'Minus_s', 'Plus_s', 'x', 'Minus_v', 'y', 'y', 'x', 'y', 'x', 'x']
	GF Index	['1.2', '3.4', '5.6', '7.8', '0.4', '9.10', '11.12', '0.7', '13.14', '0.9', '0.10', '0.11', '0.12', '0.13', '0.14']

The GF toolbox allows GF individuals to be represented in a tree-structure manner. This includes nodes in the tree, where the GF toolbox presents the results within circles and levels to build a balance tree. Figure II illustrates how each GF string in

Table III is represented by tree-structure models.



(a) Experiment 1 (b) Experiment 2

(c) Experiment 3 (d) Experiment 4

Fig. 2. Best GF Individuals found in the experiments

Table IV shows the performance results of GF versus the GA-SVM found in [16]. The accuracy comparison is found in Table II, using the SVM classifier as the measure of highest accuracy. GF outdoes all the above four types of kernels used in hybrid GA-SVM.

TABLE IV. A COMPARISON OF GF ACCURACY AGAINST [16] RESULTS

Kernel Type	***Accuracy***
Linear	0.9807
Polynomial	0.9244
Sigmoid	0.9649
RBF	0.9807
GF	**98.85**

VI. CONCLUSION

This paper introduced a novel method to classify benign and malignant breast tumors within the UCI Breast Cancer Wisconsin Dataset. The proposed algorithm initially ran with 4 different experiments with different parameters of the GF and SVM classifiers. The algorithm ran the experiments with a different number of generations, population, parameters, and GF individuals' lengths. The best result was obtained by using 5 cross-validations and 20 GF chromosome lengths without applying a crossover operator. The results were obtained by the average of 5 runs for the produced GF-SVM kernels using a mutation operator in half the population size.

Generally, GF has advantages in terms of producing accurate classifiers. From the results of the experiments, it is demonstrated that GF will have a wide range of application prospects in classification domains.

REFERENCES

[1] M. Jawad Hashim, Fatima A. Al-Shamsi, Noura A. Al-Marzooqi, Sarah S. Al-Qasemi, Ali H. Mokdad, Gulfaraz Khan. Burden of Breast Cancer in the Arab World: Findings from Global Burden of Disease, 2016. Journal of Epidemiology and Global Health. Vol. 8, Issue 1-2, Pages 54 – 58, December 2018.
[2] C DeSantis, J Ma, L Bryan, and A Jemal, Breast cancer statistics, CA Cancer J Clin, Vol. 64, 2014, pp. 52-62. , 2013. https://doi.org/10.3322/caac.21203
[3] Cortes, C. & Vapnik, V. Support-Vector Networks. Machine Learning 20: 273,1995. https://doi.org/10.1023/A:1022627411411
[4] Na Liu ,Jiang Shen, Man Xu, Dan Gan, Er-Shi Qi, and Bo Gao. Improved Cost-Sensitive Support Vector Machine Classifier for Breast Cancer Diagnosis. Advancements in Mathematical Methods for Pattern Recognition and its Applications. vol. 2018. https://doi.org/10.1155/2018/3875082

[5] T. Nadira, Z. Rustam . Classification of cancer data using support vector machines with features selection method based on global artificial bee colony. AIP Conference Proceedings. 2018. https://doi.org/10.1063/1.5064202
[6] Haifeng Wanga, Bichen Zhenga, Sang Won, Yoona, Hoo SangKob. A support vector machine-based ensemble algorithm for breast cancer diagnosis. European Journal of Operational Research. Vol. 267, Issue 2, Pages 687-699. 2018. https://doi.org/10.1016/j.ejor.2017.12.001
[7] Huang, H., Feng, X., Zhou, S. *et al.* A new fruit fly optimization algorithm enhanced support vector machine for diagnosis of breast cancer based on high-level features. *BMC Bioinformatics* **20,** 290. 2019. https://doi.org/10.1186/s12859-019-2771-z
[8] Xu, Hongyan, Chen, Ting, Lv, Junmin, Guo, Jin. A combined parallel genetic algorithm and support vector machine model for breast cancer detection. Journal of Computational Methods in Sciences and Engineering, vol. 16, no. 4, pp. 773-785, 2016. DOI: 10.3233/JCM-160690.
[9] Mohammed Ngadi, NASSIH B, Hanaa Hachimi, Aouatif Amine. Genetic algorithms combined with SVM for breast cancer diagnosis. International Workshop in Optimization and Applications. WOA2016.
[10] Zhao, Tianming. "Breast Cancer Diagnosis via the hybrid of genetic algorithm and support vector machine." 2018.
[11] Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF (2017) SVM and SVM Ensembles in Breast Cancer Prediction. PLOS ONE 12(1): e0161501. https://doi.org/10.1371/journal.pone.0161501
[12] K.Menaka, S.Karpagavalli. Breast Cancer Classification using Support Vector Machine and Genetic Programming. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 1, Issue 7, 2013.
[13] Badal Soni, Angshuman Bora, Arpita Ghosh, and Anji Reddy. FSVM: A Novel Classification Technique for Breast Cancer Diagnosis. International Journal of Innovative Technology and Exploring Engineering. Vol. 8, Issue 12, 2019.
[14] M. Mezher, M Abbod, Genetic Folding: A New Class of Evolutionary Algorithm. AI-2010 Thirtieth SGAI International Conference on Artificial Intelligence Cambridge, England 14-16 2010.
[15] Mohd Mezher. GFLIB: an Open Source Library for Genetic Folding Solving Optimization Problems. Artificial Intelligence Advances.Volume 01, Issue 01, April 2019.
[16] Tianming Zhao. Breast Cancer Diagnosis via the hybrid of genetic algorithm and support vector machine. ANU Bio-inspired Computing conference. Australia, 2018.
[17] M. Mezher, M Abbod. Genetic folding: Analyzing the mercer's kernels effect in support vector machine using genetic folding. World Academy of Science, Engineering and Technology 5 (3), 1342 – 1347.

