International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

SVM and KNN Based SGO Feature Selection Algorithm for Breast Cancer Diagnosis

P. Srihari, D. Lalitha Bhaskari

Abstract*: *In diagnosis and prediction systems, algorithms working on datasets with a high number of dimensions tend to take more time than those with fewer dimensions. Feature subset selection algorithms enhance the efficiency of Machine Learning algorithms in prediction problems by selecting a subset of the total features and thus pruning redundancy and noise. In this article, such a feature subset selection method is proposed and implemented to diagnose breast cancer using Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) algorithms. This feature selection algorithm is based on Social Group Optimization (SGO) an evolutionary algorithm. Higher accuracy in diagnosing breast cancer is achieved using our proposed model when compared to other feature selection-based Machine Learning algorithms.

Keywords: Subset Selection, Breast Cancer, Prediction, SGO, PSO, GA, CART, SVM, KNN

I. INTRODUCTION

Though breast cancer is a lethal disease, Pre-emptive Diagnosis and detection of malignancy can save many lives [1]. The primary step in detecting the malignancy in breast cancer is the classification and categorization of tumours. Tumours classified as malignant are more potentially hazardous than those which are benign. Diagnosing the tumour in the preliminary stages requires a more accurate, sophisticated and trustworthy diagnosis procedure that enables the physicians to differentiate benign and malignant tumours [2].

Computer-based solutions offer reliable and accurate information about the breast mass to the physicians than the physical interventions. These Computer-based tools providing accurate and precise diagnosis recommendations are known as Medical Diagnostic Decision Support (MDDS) systems [3].

The main issue with these MDDS systems is that they tend to have a multitude of features. The number of features can be redundant in nature and may also lead to the spike in the prediction costs and might also reflect in a very low learning precision [5,6,7]. There are a few methods suggested in the literature that provides a solution to the problem of redundant or irrelevant features. One such method is the feature selection method [8]. The selection of relevant features and choice of choosing a subset of features from the original can overcome the aforementioned persisting problem. This feature selection procedure can provide the most useful information to the user in a given dataset [9]. This process of feature selection has multiple pros associated, includes visualisation of data and decrease in the overall need for storage, processing, training and utilization times [10]. This feature selection will greatly reduce the problem called as "curse of dimensionality" [10].

Feature selection is implemented in three ways, the wrapper approach, Embedded approach and filter approach. The feature selection subset is randomly generated and evaluated for its performance by an ML algorithm, this type of approach is known as Wrapper approach. A separate scoring mechanism is used for a given subset instead of ML algorithm by Filter approach. During the training process, itself subsets are selected, this type of approach is known as Embedded approach. Wrapper approach by far is a simple approach out of the two.

Whereas, the filter approach comparatively has more cons [10]. Here, the classifier engine and the process of selecting the probable best subsets are independent of each other [10]. This issue might affect the accuracy and prediction of the classification algorithm [10]. The poor performance of the filter approach is due to the dependence of the subset selection algorithm and the similarity between data [10]. The other two approaches do not face this issue.

II. FEATURE SUBSET SELECTION TECHNIQUES

Tuba Kiyan [9] et al. attempt to implement statistical neural network for breast cancer data set. A comparison between statistical neural network and multilayer perceptron network is done on WBCD database. For classification radial basic function (RBF), General Regression Neural Network (GRNN) and Probabilistic Neural Network (PNN) were used. In RBF, PNN GRNN & MLP the over performance achieved were 96.18%, 97%, 98.8% and 95.74% respectively. Paulin and Sunthakumaran[10] designed a model implementing Back Propagation Neural Network (BPNN) and achieved 99.28% accruing with Leveaerg-Marquardt algorithm. Median filter is used for Pre-processing and normalization of data using min-max technique [10]. Karabatak and Inus et al. [11] suggested an expert model for breast cancer detection with high accuracy. For diversion reduction they have used association rules. In their analysis, AR1 and AR2 are developed for feature reduction from the original datasets. Out of 9 features AR1 and AR2 reduced to 1 and 5 features respectively.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
P Sri Hari*, Assistant Professor, Dept of IT, GMRIT, Rajam, India, Email: sriharipasala@gmail.com
D. Lalitha Bhaskari, Dept of CS & SE, Andhra University College of Engineering (A), Andhra University, Visakhapatnam, India, Email: lalithabhaskari@yahoo.co.in

By applying 3-fold cross validation they have achieved 95.6% and 95.2% with AR1 and AR2 respectively [11]. Ahemad and Ahemad et al.

[12] used three classification methods such as RBF, PNN with WBC data set. They have proved that among there with accuracy of 96.34% PNN shows better result and MLP is 96.32% [12]. Jhajharia et al.[13] used diversion reduction technique PCA for feature extraction from the data sets. Then they applied it to feed-forward network for classification. A good performance is achieved by dividing the dataset into training and test data [13]. Nayak et al.[14] implemented adaptive resource theory (ART) network for data classification. A comparative study is done with PSO-MLP and PSO-BBO algorithm. Result analysis proved that ART performed better with respect to other two classifiers [14]. Nilashi et al. [15] introduced a new methodology combining fuzzy logic with knowledge base system with breast cancer data set. They have used PCA for data reduction. For generation of fuzzy rules CART is used for knowledge based system. For classification fuzzy rule based reasoning method is introduced [15]. Virmani et al.[16] presented a comparative analysis on SV classifier. They designed PCA with SVM for breast cancer data classification. Result analysis claims PCA with SVM performed well then SVM [16]. Jihong Liu et al.[17] introduced a methodology by combining two classifiers such as SVM and BPNN with PCA a features extraction algorithm for breast cancer images. SVM proved the best classifier as per the result achieved [17]. Gouda I. Salama et al. [18] three different classifiers such as C 4.5, SVM and MSP along PCA with 3 breast cancer datasets [18]. Kavakiotis, I.,et al. [19] proposed a hybrid method for feature selection using PCA and C 4.5 with heart disease datasets. The classification result was quick satisfactory [19].

III. SGO

Social Group Optimization [SGO] is an optimization algorithm which consists of two parts namely the improving phase and the acquiring phase [20]. The best person in the group influences each person in the group and enhances their knowledge, this phase is known as the improving phase. This best person in the group contains the highest knowledge in problem-solving, as he disseminates knowledge to each person, their existing knowledge gets updated, this phase is known as acquiring phase.

Assume $X_{j\ =\ 1,\ 2,\ 3,\ldots N}$ be the considered persons of a particular social group, i.e., the considered social group will have a total of N persons and the individual person X_j is defined by $X_j = (x_{j1},\ x_{j2},\ x_{j3},\ldots,\ x_{j\ D})$, Here D refers to the total number of traits that a particular individual has, this will show the dimensions of that particular individual. Also, $f_{j\ =\ 1,\ 2,\ldots N}$ represents the fitness values of those persons mentioned above.

Improving phase

The G_{best} value person shares his knowledge with all other persons in the group, this sharing will improve the knowledge of others in the group [20].

Hence, $g_{best,}\ g = \max(f_{i,})\ \{i = 1, 2,\ldots N\}$ at generation g for solving maximization problem.

There are two phases in the SGO algorithm, the first phase, improving phase, is where each particular person in the group obtains knowledge from the best person in the group known as g_{best}. To update each person the following function is computed:

For i = 1 : N

  For j=1:D

    $X_{newij} = (X_{oldij} * c) + (g_{bestj} - X_{oldi\ j}) * r$

  End for

End for

Here the value of r is a random number, r~ U(0, 1)

X_{new} gets accepted if and only if it has a better fitness than X_{old}

where c is known as self-introspection parameter. Its value can be set from 0 < c < 1.

Acquiring phase

A random person in the group will interact with the best person in that group and also communicates with the other persons in that group which acquires him knowledge. Acquiring of knowledge is possible if the person has lesser knowledge than the best person in the group.

The below pseudo code explains the acquiring phase.

$g_{best} = \min\{ f(X_i), i = 1, 2,\ldots N\}$

For i = 1 : N

  Randomly select one person X_r, where i = r

  If $f(X_i) < f(X_r)$

    For j = 1 : D

      $X_{newi,j} = X_{oldi,j} + r_1 * X_{i,j} - X_{r,j} + r_2 * (g_{bestj} - X_{i,j})$

    End for

  Else

For j = 1 : D

  $X_{newi,:} = X_{oldi,:} + r_1 * X_{r,:} - X_{i,:} + r2 * (g_{bestj} - X_{i\ j})$

End for

3. SGO based Subset Selection (SGOss)

Satapathi et al. [20] carried put various 10 different experiments on 30 bench marks standard functions comparing evolutionary algorithms such as Genetic algorithm, TLBO, ACO, PSO etc. From the results of their study it can be observed that SGO has outclassed all other evolutionary algorithms in terms of performance and costs. Hence, we chose to use the SGO algorithm as a base to develop a new feature subset selection algorithm.

This is a wrapper approach, i.e., machine learning algorithms such as KNN and SVM are used to evaluate the subsets. Here, to evaluate the subsets, accuracy of SVM and KNN classifiers are used. This algorithm consists of four phases.

First Phase:

In the first phase, the WBCD dataset [21] is normalised to range of values between 0 and 1. This removes the disparities in ranges of various attributes.

Second Phase:

Once the dataset is pre-processed, in the second phase, a subset selection parameter is chosen to determine the no of subsets we would like to be generated. As the total size of the dataset is 569 x 32 after pre-processing Label, S.no and class are removed and the dimensions are 569 x 30. A Subset selection parameter value of 2 will generate 435 subset combinations. Of these, 435 combinations, a random combination without a repetition is selected. Also, iteration count is set, to specify how many different combinations need to be tried.

Third Phase:

On these selected subsets, a maximum fitness/objective function is applied. Here accuracy from SVM and KNN algorithms is used as a fitness function. To apply the fitness function, the data set is split into 60-40 ratio for training and test split and a 10-fold cross validation is used.

Once the fitness is calculated, the g_{best} is updated using the improving phase and acquiring phase of SGO algorithm. The algorithm is terminated if the iteration count is reached. The following flowchart describes the above process.

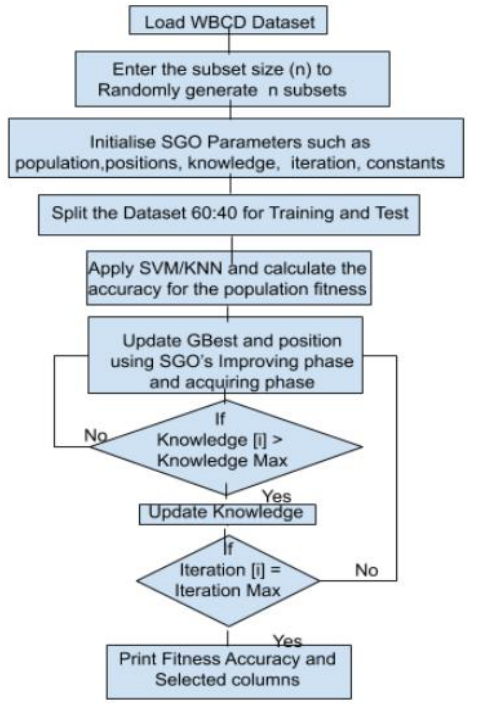

Figure 1. Flowchart SGO Feature Selection Algorithm using SVM/KNN

IV. RESULTS

Our proposed approach the SGO based Feature selection using SVM/ KNN is evaluated on two different modes. Many of the existing models used 80:20 test split for the accuracy, though it can contribute to higher accuracy as the model has more training sample. The industry standard train test split is at 60:40, and also cross validation of 10. Hence, we used both 80:20 and 60:40 split with cross validation for our comparison.

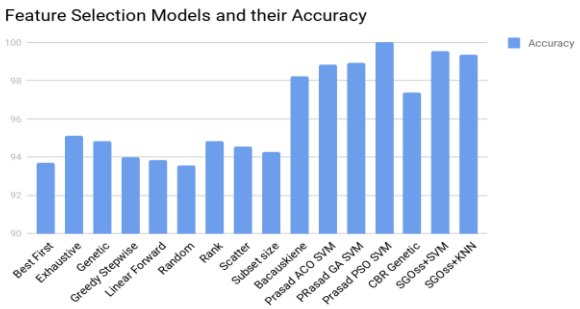

Figure 2. Feature Selection models and their accuracy.

The Accuracy of our proposed model with the existing models is shown in figure2 and the corresponding attributes selection with accuracy is shown in table1.

Table1. Subset Selection Models with No.of selected attributes and their accuracy.

Method	Attributes	Accuracy% (80:20)
Best First	7	93.7
Exhaustive	6	95.13
Genetic	6	94.84
Greedy Stepwise	7	93.99
Linear Forward Selection	7	93.84
Random	7	93.56
Rank	9	94.84
Scatter	7	94.56
Subset size forward selection	7	94.27
Bacauskiene	10	98.24
Prasad ACO SVM	20	98.83
Prasad GA SVM	7	98.95
Prasad PSO SVM	18	100
CBR Genetic	12	97.37
SGOss+SVM	**2**	**99.56**
SGOss+KNN	**2**	**99.38**

Apart from the above, evaluations. Accuracy based on 60:40 test split is also made and several comparisons are made based on number of iterations. The table 2 below shows the results for 100 iteration and the subsets selected.

Table 2. Subsets selected and their Accuracy for 100 iterations on 60:40 split.

	SVM		KNN	
No.of Attributes	Selection Value	Accuracy%	Selection Value	Accuracy%
2	**11, 10**	**97.36**	**3, 26**	**96.05**
3	**26, 12, 21**	**96.92**	**4, 8, 28**	**96.49**
4	**26, 9, 10, 15**	**97.36**	**11, 16, 10, 27**	**96.05**
5	**16, 9, 5, 3, 6**	**97.36**	**25, 9, 29, 2, 26**	**96.92**

Our results suggest that SVM has an overall better accuracy when compared with KNN algorithm for different subset sizes and train test splits.

V. CONCLUSION

SVM and KNN are used with SGO algorithm to select few features of WBCD dataset as to reduce the overall running time of the prediction systems. We have evaluated our proposed model with the existing models using 80:20 split and observed that our model attains an accuracy of 99.56% with only 2 subsets. Though there exists another model which produces 100% accuracy it requires 18 features out of the 30 features. Apart from the above evaluation we have also tested the model on a 60:40 test split with 10-fold cross validation and observed that SVM with SGOss algorithm produced an accuracy of 97.36 with 2, 4 or 5 subsets. When compared with KNN, SVM had an overall better performance in feature subset selection.

REFERENCES

1. I. Harirchi, et al., "Breast cancer in Iran: a review of 903 case records," Public Health, 2000. 114(2): p. 143-145.
2. T. Subashini, V. Ramalingam, and S. Palanivel, "Breast mass classification based on cytological patterns using RBFNN and SVM," Expert Systems with Applications, 2009. 36(3): p. 5284-5290.
3. R.A. Miller, "Medical diagnostic decision support systems - past, present, and future," Journal of the American Medical Informatics Association, 1994. 1(1): p. 8. World Academy of Science, Engineering and Technology International Journal of Biomedical and Biological Engineering Vol:5, No:5, 2011
4. J. Han, and M. Kamber, "Data mining: concepts and techniques," 2006: Morgan Kaufmann.
5. R. Kohavi, and G.H. John, "Wrappers for feature subset selection," Artificial intelligence, 1997. 97(1-2): p. 273-324.
6. Y. Yuling, "A Feature Selection Method for Online Hybrid Data Based on Fuzzy-rough Techniques," 2009: IEEE.
7. N. Abe, et al., "A divergence criterion for classifier-independent feature selection," Advances in Pattern Recognition, 2000: p. 668-676.
8. M. Dash, and H. Liu, "Feature selection for classification," Intelligent data analysis, 1997. 1(3): p. 131-156.
9. Kiyan, T., & Yildirim, T. (2004). Breast cancer diagnosis using statistical neural networks. IUJournal of Electrical & Electronics Engineering, 4(2), 1149-1153.
10. Paulin, F., & Santhakumaran, A. (2011). Classification of breast cancer by comparing back propagation training algorithms. International Journal on Computer Science and Engineering, 3(1), 327–332.
11. Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications, 36(2), 3465–3469.
12. Azar, A. T., & El-Said, S. A. (2014). Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Computing and Applications,24(5), 1163–1177.
13. Jhajharia, S., Varshney, H. K., Verma, S., & Kumar, R. (2016). A neural network based breast cancer prognosis model with pca processed features. In: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1896–1901.
14. Nayak, T., Dash, T., Rao, D. C., & Sahu, P. K. (2016, August). Evolutionary neural networks versus adaptive resonance theory net for breast cancer diagnosis. In Proceedings of the International Conference on Informatics and Analytics (p. 97). ACM
15. [15]Nilashi, M., Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics and Informatics, 34(4), 133–144.
16. [16]Virmani, J., Dey, N., & Kumar, V. (2016). PCAPNN and PCA-SVM based CAD systems for breast density classification. In Applications of intelligent optimization in biology and medicine(pp. 159-180). Springer, Cham.
17. Liu, J., & Ma, W. (2008, January). An effective recognition method of breast cancer based on PCA and SVM algorithm. In International Conference on Medical Biometrics (pp. 57-64). Springer, Berlin, Heidelberg.
18. Salama, G. I., Abdelhalim, M., & Zeid, M. A. E. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC), 32(569), 2.
19. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.
20. Satapathy SC, Naik A (2016) Social group optimization (SGO): anew population evolutionary optimization technique. J ComplexIntell Syst. doi:10.1007/s40747-016-0022-8.
21. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

AUTHORS PROFILE

P. Srihari is a research scholar in Andhra University under the supervision of Prof. D. Lalitha Bhaskari in Computer Science and Systems Engineering. He received his M.Tech (CST) fr**om GITAM University and presently working as Assistant** Professor in IT Department of GMRIT. His research areas include Machine learning and Image processing.

D. Lalitha Bhaskari is a Professor in the department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, India. She is guiding more than 15 PhD Scholars from various institutes. Her main research interest includes Network Security, Image Processing, Pattern Recognition, steganography and Digital Watermarking. Prof. D. Lalitha Bhaskari is a member of IEEE, IJSCI,CSI and Associate Member of Institute of Engineers