Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 16 – 23

Procedia Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

# A Novel Feature Selection Technique for Improved Survivability Diagnosis of Breast Cancer

S.Sasikala[a], Dr.S.Appavu alias Balamurugan[b] and Dr.S.Geetha[c*]

[a]*Research Scholar, Anna university, Tamil Nadu, India, nithilannsasikala@yahoo.co.in*

[b]*Professor and Head, K.L.N. College of Information Technology, Tamil Nadu, India, app_s@yahoo.com*

[c] *Professor, School of Computing Science and Engineering, VIT University, Tamil Nadu, India, geetha.s@vit.ac.in*

**Abstract**

In this paper we propose a novel Shapely Value Embedded Genetic Algorithm, called as SVEGA that improves the breast cancer diagnosis accuracy that selects the gene subset from the high dimensional gene data. Particularly, the embedded Shapely Value includes two memetic operators namely "include" and "remove" features (or genes) to realize the genetic algorithm (GA) solution. The method is ranking the genes according to its capability to differentiate the classes. The method selects the genes that can maximize the capability to discriminate between different classes. Thus, the dimensionality of data features is reduced and the classification accuracy rate is improved. Four classifiers such as Support vector machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN) and J48 are used on the breast cancer dataset from the Kent ridge biomedical repository to classify between the normal and abnormal tissues and to diagnose as benign and malignant tumours. The obtained classification accuracy demonstrates that the proposed method contributes to the superior diagnosis of breast cancer than the existing methods.



Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

*Keywords:* Medical Data mining, Breast Cancer diagnosis, Bio medical classification, Feature selection, Shapley values, Genetic Algorithm

## 1. Introduction

Breast cancer is the most common cancer affecting women all over India and accounts for about 25% to 31% of all types of cancers in women in Indian cities. It is highly warning to observe that the average age by which an individual develops breast cancer has shifted from 50 - 70 years to 30 - 50 years; and cancers in the young is inclined to be more aggressive. As per the reports of World Health Organization (WHO), for the year 2013, it is estimated that around 70218 women died in India due to breast cancer, which is more than any other country in the world (second: China - 47984 deaths and third: US - 43909 deaths ). A positive hope is that women with a higher than



Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)
doi:10.1016/j.procs.2015.04.005

average risk of developing breast cancer, like one being family history, may be suggested for screening and genetic testing periodically for the condition, as a pro-active measure. Approximately 20% of the cancers detected in a particular year even though missed at the initial screening become clinically evident in the period before the next screen (interval cancers).

The core challenge with the DNA micro-array structure is that the dataset contains massive collection of data. For example, the breast cancer micro-array (Jinyan & Huiqing, 2002) contains the samples of few patients (usually less than 100) but each of the samples includes more than 20,000 genes. Among these the cancer signature bearing genes are less and those markers are present at varied positions in each of the patient's sample. Hence identifying the genes and inferring the knowledge from this massive quantity of varying genes really enforces challenge to the diagnosis process. Among the thousands of genes, the potential ones that denote the presence of cancer, could be identified and used for further knowledge inference. This necessitates for the feature selection process to be applied as a pre-processing activity. The conventional feature selection steps produce sub-optimal results since they are not targeted at gene selection on a micro-array dataset. They have been designed for small and medium scale datasets. Motivated by these factors, this paper aims at developing a custom-built feature selection algorithm that reduces the genes to a considerable amount and minimizes the complexity on further processing, pre and post diagnosis.

Feature selection, as a pre-processing step to machine learning, is prominent and effective in dimensionality reduction, by removing irrelevant and redundant data, increasing learning accuracy, and improving result comprehensibility. Feature selection methods (Jazzar & Muhammad, 2013; Shen,Diao & Su, 2011; Han & Kamber, 2006) tend to identify the features most relevant for classification and can be broadly categorized as either subset selection methods or ranking methods. The former type returns a subset of the original set of features which are considered to be the most important for classification (Bolo n-Canedo,Sanchez-Marono & Alonso-Betanzos , 2012). Ranking methods sort the features according to their usefulness in the classification task.

## 2. Proposed Method –SVEGA- Shapley Value –Embedded Genetic Algorithm

Shapley Value Analysis has been proved to be a promising strategy for feature selection process. Shapley Value Analysis (SVA)( Jeffery et al ., 2006; Bu Hualonga & Jingb X, 2011) is a game theory based technique for causal function localization that addresses the issue in describing and calculating the contributions made by the interactions among the group of elements in the breast-cancer micro-array data set with multiple features and their corresponding performance scores.

In this section, the proposed memetic algorithm, particuarly, Shapley Value Embedded GA (SVEGA) is outlined. At the beginning of the SVEGA search, the population for GA (Senthamarai Kannan & Ramaraj, 2010) is initialized randomly where each chromosome in the pool encodes a candidate feature subset. In this work, each chromosome is built of a binary string whose length equals the total number of features in the dataset of interest. In binary encoding, a bit of value '1' ('0') indicates that the respective feature is selected (omitted). The objective function for calculating the fitness of each chromosome is then obtained as follows:

$$\text{Fitness}(c) = \text{Obj\_Fun}(SFc) \quad (1)$$

Where SFc denotes the Selected Feature subset encoded by a given chromosome c, and the objective function for feature selection Obj_Fun(SFc) calculates the contribution of the given feature subset SFc. We use the classification accuracy and number of features generated as the metrics in our Obj_Fun (SFc). The former one requires maximization and the latter one is to be minimized. i.e. maximum accuracy and minimum number of features. In case of two chromosomes having same fitness value, the chromosome with smaller number of selected features is given higher priority of surviving and is moved on to the next generation. This is recommended in a feature classification problem, where a subset of features with fewer features giving higher classification accuracy is preferred over a subset of features with more features giving lower or equal classification accuracy. The Pseudo code for Shapley Value Embedded Genetic Algorithm (SVEGA) for feature selection is shown in Figure 1.

The pseudo-code of the algorithm is given below:

Shapley Value Embedded Genetic Algorithm (SVEGA)
BEGIN
Population Initialization:
An initial population of size 50 encoded with binary string is randomly generated. (A gene value of '1' means, the feature at that position is selected and a value of '0' means, the feature at that position is omitted)
While(*not reached the convergence point or computational cost is not over*)
Fitness Evaluation
The fitness value of all feature subsets in the population is evaluated according to $Obj\_Fun(SF_c)$.
Selection:
The elite chromosome $c_e$ is selected and subjected to Shapley Value based memetic operations.
Lamarckian learning
The elite chromosome $c_e$ is replaced with improved new chromosome $c_e^{''}$ by Lamarckian Learning process.
Evolutionary Operations:
The evolutionary operations like linear ranking selection, restrictive crossover and mutation operator with elitism.
End While
END

Figure 1. Proposed Shapley Value Embedded Genetic Algorithm (SVEGA) for feature selection

***Include* Operator**
**BEGIN**
(1) Rank the features in $R$ in decreasing order of their *Shapley* values.
(2) Select a feature $R_i$ in $R$ by linear ranking selection in such a way that a feature with larger *Shapley value* of a feature in $R$ is more likely to be selected.
(3) Add $R_i$ to $Q$.
**END**

Figure 2. pseudo code of memetic operators: Include Operator

***Remove* Operator**
**BEGIN**
(1) Rank the features in $Q$ in decreasing order of Shapley value.
(2) Select a feature $Q_i$ in $Q$ by linear ranking selection in such a way that a feature with larger *Shapley value* of a feature in $Q$ is more likely to be selected.
(3) Eliminate all the features in $Q-\{Q_i\}$.
**END**

Figure 3. pseudo code of memetic operators: Remove Operator

In each of the GA generation, the elite chromosome, i.e., the chromosome having the best fitness value is selected and subjected to Shapley value based memetic operators as a part of the Lamarckian learning process (Krasnogor, 2002; Ong & Keane, 2004). The Lamarckian learning (Minseok Seo & Sejong Oh 2012) brings improvement in the result by placing the locally improved individual genes back into the population pool so that they acquire the reproductive opportunities. We define two memetic operators in the SVEGA, namely an Include operator which includes/adds a feature to the elite chromosome, and a Remove operator which removes/omits the existing features from the elite chromosome. The key issue is deciding which features to include and which ones to omit. Preferably, the features to be removed will be the ones which provide the least contribution when considered as a whole set and the ones which provide highest contribution must be included into the solution feature subset. This characteristic has to be brought in the existing GA paradigm. This requirement is fulfilled by the use of Shapley value concept.

After executing the above given Lamarckian learning process over the elite chromosome, the GA population then goes through the typical evolutionary operations like linear ranking selection (Baker, 1985), restrictive crossover, and mutation operators with elitism (Moretti et al., 2008) Since we had a prior knowledge on the optimum number of features for certain datasets, we allowed the integration of such information into our proposed SVEGA by limiting the number of '1' bits in each of the chromosome to a maximum of $'m'$ ($'m'$ is safely chosen to greater

than the optimum number of features) during the evolutionary search process. To facilitate this aspect, we employed restrictive crossover operator and mutation (Zhu et al., 2006) rather than the conventional evolutionary operators of GA, so that the number of '1' bits occurring in each chromosome does not break the constraint imposed by the prior knowledge on $'m'$ during the search. The function of the Include operator is to identify and select the feature with maximum Shapley score when measured in coalition, from set $R$ and pushes it to the set $Q$ .On the other hand, the Remove operator serves to identify and select the features with minimum contribution score and deletes from set $Q$ and moves that into the set $R$ . The pseudo code of these memetic operators is outlined in Figure 2 and Figure 3. The pseudo code of the Shapley value embedded memetic operation executed on the elite chromosome of each of the GA generation is outlined in Figure 4.

## 3. Experimental Results and Discussion

### *3.1. Experimental Scenario*

The proposed approach has been evaluated by experiments on breast cancer micro array from the Kent ridge biomedical repository (Jinyan & Huiqing,2002 ).The training data contains 78 patient samples, 34 of which are from

**Shapley Value Based Memetic Operation**
BEGIN
Select the elite chromosome $c_e$ to undergo memetic operations.
For $j = 1 \; to \; L^2$
Generate a unique random pair of values $\{i, r\}$ where $0 \le i, r \le L$ .
Apply $'i'$ times *Include* on the elite chromosome $c_e$ and generate a new chromosome $c_e^{'}$.
Apply $'r'$ times *Remove* on $c_e^{'}$ and generate a new chromosome $c_e^{''}$ .
Calculate the fitness of new modified chromosome $c_e^{''}$ based on $Obj\_Fun(SF_c)$.
If $c_e^{''}$ is better than $c_e$ either on fitness value or the number of features
Replace the genotype $c_e$ with $c_e^{''}$ and stop applying the memetic operation.
End If
End For
END

Figure 4. Shapley Value Based Memetic Operation

patients who had developed distance metastases within 5 years (labeled as "relapse"), the rest 44 samples are from patients who remained healthy from the disease after their initial diagnosis for interval of at least 5 years (labeled as "non-relapse"). Correspondingly, there are 12 relapse and 7 non-relapse samples in the testing data set. The number of genes is 24481.

The Table 1 summarizes the classification performance in terms of average accuracy and running time for the specified conventional classifiers on whole dataset before applying the proposed SVEGA method and on the reduced data set after applying the proposed SVEGA method.

### *3.2. Classification Results using Proposed SVEGA Framework*

In Table 1 the evaluation measures are recorded on both unprocessed (original) features and reduced features by the proposed SVEGA method and compared with other methods as specified in references (Feng et al., 2008 ; Minseok Seo & Sejong Oh 2012).From the measures the proposed method SVEGA stands superior when compared with other methods. The best measure in each table is highlighted in bold typeface. Figure 6 shows the dimensions

of features selected from the original dataset by existing and proposed feature selection. From the Figure 5 and Figure 6, it has been observed that SVEGA method works well than the other existing methods and are best with the property of selecting minimum number of features for classifying normal patients (samples).

*3.3. Performance Metric*

For measuring the performance of the proposed system, we use the following metrics. We present them in view of binary class problem which give two discrete outputs positive class and negative class. In binary classification, for a given classifier and instance, we have four possible outcomes.

True Positive (TP) – Positive instances correctly classified as positive outputs
True Negative (TN) – Negative instances correctly classified as negative outputs
False Positive (FP) – Negative instances wrongly classified as positive outputs
False Negative (FN) – Positive instances wrongly classified as negative outputs

$$True\ Positive\ Rate\ (TPR) = \frac{Postives\ correctly\ classified}{Total\ number\ of\ positives} = \frac{TP}{TP+FN} \tag{2}$$

$$False\ Positive\ Rate\ (FPR) = \frac{Negatives\ correctly\ classified}{Total\ number\ of\ negatives} = \frac{FP}{FP+TN} \tag{3}$$

$$Classification\ Accuracy\ is\ acc = \frac{TP+TN}{TP+FP+TN+FN} \tag{4}$$

$$Sensitivity(SN) : SN = TPR \tag{5}$$

$$Specificity(SP) : SP = 1 - FPR \tag{6}$$

$$Precision = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ False\ Positives} \tag{7}$$

$$Recall = Senstivity = \frac{Postives\ correctly\ classified}{Total\ number\ of\ positives} = \frac{TP}{TP+FN} \tag{8}$$

Table 1: Classifiers performance on the features selected by the existing systems and proposed SVEGA method on Breast cancer Dataset

| Classifiers | Observations | None | Bu et al. 2011 [9] | Feng et al. 2008 [17] | Senthamarai et al. 2010 [10] | Proposed SVEGA System |
|---|---|---|---|---|---|---|
| ***K-NN Classifier*** | Accuracy | 62.88 | 69.56 | 75.63 | 79.57 | **82.47** |
| | No. of features selected | 24482 | 15478 | 4236 | 183 | **41** |
| | Precision | 0.318 | 0.489 | 0.812 | 0.818 | **0.894** |
| | Sensitivity | 0.682 | 0.546 | 0.712 | 0.818 | **0.887** |
| | Specificity | 0.680 | 0.532 | 0.582 | 0.818 | **0.887** |
| | F-measure | 0.681 | 0.479 | 0.499 | 0.816 | **0.886** |
| | Running Time (sec) | 626 | 612 | 526 | 432 | **411** |
| ***NB Classifier*** | Accuracy | 54.63 | 64.56 | 82.84 | 92.56 | **88.50** |
| | No. of features selected | 24482 | 15478 | 4236 | 183 | **41** |
| | Precision | 0.756 | 0.489 | 0.512 | 0.592 | **0.835** |
| | Sensitivity | 0.546 | 0.546 | 0.612 | 0.716 | **0.835** |
| | Specificity | 0.546 | 0.532 | 0.582 | 0.642 | **0.835** |
| | F-measure | 0.407 | 0.479 | 0.499 | 0.624 | **0.835** |
| | Running Time (sec) | 714 | 654 | 631 | 542 | **436** |
| ***SVM Classifier*** | Accuracy | 68.04 | 72.34 | 83.86 | 90.23 | **91.75** |
| | No. of features selected | 24482 | 15478 | 4236 | 183 | **41** |
| | Precision | 0.61 | 0.489 | 0.512 | 0.784 | **0.939** |
| | Sensitivity | 0.608 | 0.546 | 0.612 | 0.784 | **0.938** |
| | Specificity | 0.608 | 0.532 | 0.582 | 0.784 | **0.938** |
| | F-measure | 0.599 | 0.479 | 0.499 | 0.783 | **0.938** |
| | Running Time (sec) | 723 | 712 | 686 | 358 | **341** |
| ***J48 Classifier*** | Accuracy | 60.82 | 77.56 | 83.12 | 88.56 | **93.81** |
| | No. of features selected | 24482 | 15478 | 4236 | 183 | **41** |
| | Precision | 0.629 | 0.489 | 0.512 | 0.754 | **0.827** |
| | Sensitivity | 0.629 | 0.546 | 0.612 | 0.753 | **0.825** |
| | Specificity | 0.629 | 0.532 | 0.582 | 0.753 | **0.825** |
| | F-measure | 0.625 | 0.479 | 0.499 | 0.753 | **0.824** |
| | Running Time (sec) | 532 | 524 | 496 | 451 | **404** |

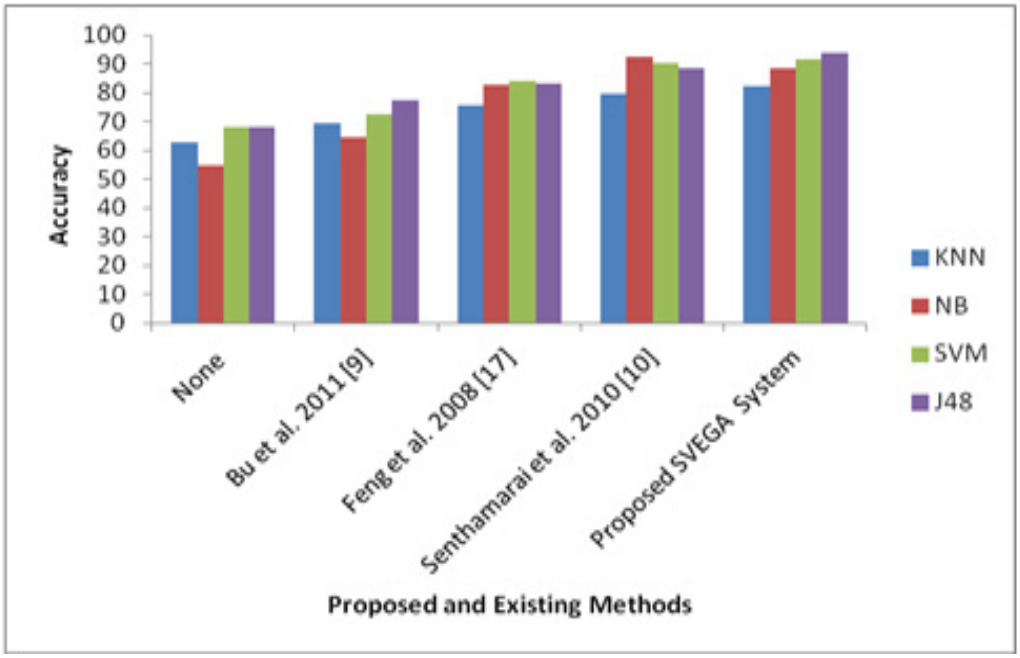


Figure 5.Classifier performance for the existing and proposed feature selection methods

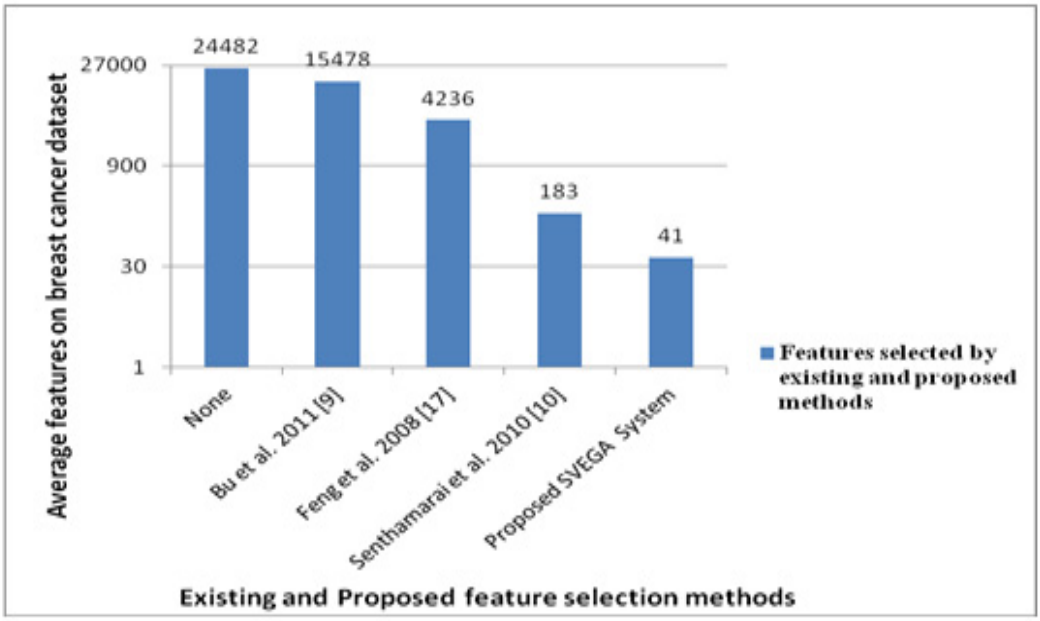


Figure 6.Average features selected by existing and proposed Feature Selection Methods

## 4. Conclusion

Feature selection combining with classification models in microarray technology play an important role in diagnosing and predicting disease, in medical research. A feature selection method SVEGA for finding the most significant features is proposed. The classification accuracy rate achieved by the proposed SVEGA method using four classifiers such as J48 is 93.81%, SVM is 91.75%, NB is 88.5% and KNN is 82.476 %.The number of features reduced from 24,481 to minimum of 6 features. Experimental results on the Breast cancer datasets clearly indicate that the proposed technique has better performance compared to the existing method.

## Acknowledgements

This work is supported in part by the University Grant Commission (UGC), New Delhi, INDIA -Major Research Project under grant no. F.No.:39-899/2010 (SR).

## References

1. Jinyan, L & Huiqing, L. Kent ridge bio-medical data set repository Retrieved April 5, 2014 from the World Wide Web: Http://datam.i2r.a-star.edu.sg/datasets/krbd.
2. Xing, E.,Jordan, M., & Karp, R.,(2001) Feature Selection for High-Dimensional Genomic Microarray Data. In: proceedings of Machine Learning (pp. 601-608). Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.
3. Jazzar, M,M., & Muhammad G.,(2013) Feature Selection Based Verification /Identification System Using Fingerprints and Palm Print. Arabian Journal for Science and Engineering, 38(4): 849-857.
4. Shen, Q, Diao R., & Su P.,(2011) Feature Selection Ensemble. In: proceedings of Computing, Springer-Verlag, (pp. 289-306).
5. Han, J, W. & Kamber, M. (2006).Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.
6. Bolo n-Canedo V,Sanchez-Marono N, Alonso-Betanzos A.,(2012) An ensemble of filters and classifiers for microarray data classification. Pattern Recognition, 45(1), 531–539.
7. Mohammad Darzi., Ali AsgharLiaei., Mahdi Hosseini., Habibollah Asghari., (2011) Feature Selection for Breast Cancer Diagnosis: A Case-Based Wrapper Approach, World Academy of Science, Engineering and Technology, 77(1).
8. Jeffery, I.B., Higgins, D.G., Culhane, A.C., (2006) Comparison and evaluation of methods for generating differentially expressed gene lists from microarray Data, BMC Bioinformatics, 7:359.
9. Bu Hualonga., & Jingb X, (2011) .Hybrid Feature Selection Mechanism based High Dimensional Date sets Reduction, Energy Procedia, 11(1): 4973-4978.
10. Senthamarai Kannan S., & Ramaraj N, (2010). A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm, Knowledge-Based Systems, 23(1): 580–585.
11. Krasnogor N., (2002). Studies on the Theory and Design Space of Memetic Algorithms, Ph.D. Thesis, Faculty of Computing, Mathematics and Engineering, University of the West of England, Bristol, U.K.
12. Ong, Y.S., & Keane, A.J.(2004).Meta-Lamarckian in Memetic Algorithm, IEEE Trans. Evolutionary Computation , 8(2):99-110.
13. Minseok Seo., & Sejong Oh., (2012). Derivation of an artificial gene to improve classification accuracy upon gene selection, Computational Biology and Chemistry , 36(1): 1–12.
14. Baker, J.E., (1985) .Adaptive Selection Methods for Genetic Algorithms, In: proceedings of International conference on Genetic Algorithm and Their Applications (pp.101-111); Holland J, H.: Adaptation in natural artificial systems. 2nd edition, MIT Press.
15. Moretti S., Danitsja van Leeuwen., Hans Gmuender., Stefano Bonassi., Joost Van Delft., Jos Kleinjans.,Fioravante Patrone., & Domenico Franco Merlo.(2008) Combining Shapley value and statistics to the analysis of gene expression data in children exposed to air pollution,BMC Bioinformatics,9:361:1-21.
16. Zhu Z., Ong YS., & Dash M.(2006) Wrapper-Filter Feature Selection Algorithm Using A Memetic Framework, IEEE Transactions on Systems, Man and Cybernetics - Part B ,10(4):392-404.
17. Feng Tan., Xuezheng Fu., Yanqing Zhang., Anu G., & Bourgeois. (2008) A genetic algorithm-based method for feature subset selection, Soft Computing, 11(1):111–120.