SVM &GA-CLUSTERING BASED FEATURE SELECTION APPROACH FOR BREAST CANCER DETECTION

Rashmi Priya[1]and Syed Wajahat Abbas Rizvi[2]

[1]Assistant Professor GD Goenka University ,India
[2] Amity University, Uttar Pradesh ,India

ABSTRACT

Mortality leading among women in developed countries is breast cancer. Breast cancer is women's second most prominent cause of cancer mortality worldwide. In recent decades, women's high prevalence of breast cancer has risen dramatically. This paper discussed several data analysis methods used to detect breast cancer early. Breast cancer diagnosis distinguishes benign and malignant breast lumps. Using data processing tools, we tackled this disease analysis. Data mining is an important step of library discovery where intelligent methods are used to detect patterns. Several clinical breast cancer studies were conducted using soft computing and machine learning techniques. Sometimes their algorithms are easier, easier, or more comprehensive than others. This research is focused on genetic programming and machine learning algorithms to reliably identify benign and malignant breast cancer. This study aimed to optimise the testing algorithm. We used genetic programming methods to choose classification machines' best features and parameter values. Data mining is an important step of library discovery where intelligent methods are used to detect patterns. We are analysing data accessible from the U.C.I. deep-learning data set in Wisconsin. In this experiment, we equate four Weka clustering strategies with genetic clustering. A comparison of results reveals that sequential minimal optimization (S.M.O.) is better than I.B.K. and B.F. Tree processes, i.e. 97.71%.

KEYWORDS

S.M.O., Breast cancer, Machine learning, Feature selection, and WEKA

1. INTRODUCTION

Breast cancer is the most prevalent non-skin cancer in women and the second-largest cause of cancer mortality in women. [1]. [1]. "Mammography usually represents thick-area breast cancer and clusters. A normal benign mass has a circular border, circumscribed and round, but malignant cancer usually has a suspected raw and fuzzy boundary. "[2],[3].

Nowadays, demand for machine learning grows until it becomes an operation. Sadly, machine learning also takes skills and is a field with very high barriers. The creation of a successful machine learning model involves many skills and experience, including pre-processing steps, feature selection, and classification processes. Using computer analysis and machine learning methods in medical fields is prevalent, as these techniques can be regarded as a great aid in decision-making processes for medical professionals. A vast variety of databases are being used for breast cancer cases that are helpful in supporting scientific and academic studies and even more in integrating previous field computer analysis and machine learning.

DOI :10.5121/ijscai.2020.9401

In certain fields and implementations, solving such a problem is based on retrieving features from the original images obtained in the physical world and organised as vectors. The processing system 's efficiency depends heavily on the correct choice of these vectors. But, in many situations, problem solving becomes almost impossible due to the excessive dimensionality of these vectors or anomalies that may exist in the results. Therefore, reducing the size of the dataset samples to a more suitable size is sometimes helpful and often necessary, even though this reduction may lead to minor information loss.

Accurate and accurate data will be gathered to help the doctor's early diagnosis and treatment of illness, both stable and malignant, using an exact model. This will save doctors time and improve efficiency. This article reflects on how benign or malignant breast cancer diagnosis is, also at an early stage. Forecast criteria are based on breast cancer symptoms. This paper's data set contains 32 attributes.

A breast cancer diagnosis may be useful in predicting the outcomes of complex diseases or recognising the molecular nature of the tumour. Many methods for examining and identifying trends of breast cancer. This paper compares empirically the utility of three classical tree classifiers designed to specifically evaluate their effects. Traditional SVM preparation normally requires Q.P. Set, and it takes longer to solve Q.P. Optimization problem, particularly for large dataset problems. SVM planning is slow, and the creation of large data sets takes time. The SMO-SVM method minimises data, is more reliable and easier to implement[4]. We suggest a hybrid SVM-GA (Support vector machines with genetic attributes) approach to achieve optimal results on Dataset breast cancer.

2. LITERATURE SURVEY

Machine learning techniques usually refer to life scientific analysis. Numerous research focused on medical diagnostic technology has been written. These studies have applied different solutions to the problem and achieved detailed classification precision[5] using an artificial cortical network to assess breast cancer therapy. They have tested their system on a limited set of data, but results suggest they understand actual survival. And al.[6] Also in breast cancer patients, a naïve bay, decision tree, and neural backpropagation network were used.

Though the results were high (about 90% accuracy), they were not appropriate because the data were split into two groups: one for more than five years of life and the other for those who died in five years. Findings became meaningless. [7] Program pick approach to usability assessment of feature selectors. This seeks a simple, coherent set of functions without losing the predictive precision dimension. Using a ranking algorithm applies confidence to characteristics. [8] Proposed hybrid GA / SVM approach using fuzzy logic to minimise the initial problem size. Identify a sub-set of balanced genes, which are then checked by SVM. [9] This analysis aimed to compare the performance of the Artificial Neural Network.

(ANN) and Vector Machine Support (SVM) for liver cancer classification. On BUPA Liver Disease Dataset, both model accuracy, durability, precision, and performance were contrasted and validated. Curved Field (A.U.C.). [10] used in tandem with heart disease modelling in mining and genetic algorithms. The proposed approach used Gini genetic mutation index statistics for interface process and crossover. They used a technique for gathering consistency functions.

3. CLASSIFICATION

Classification is one of the data mining techniques specifically used to analyse and assign a given dataset to a particular class[11]. This method is designed to remove classification errors.

Classification enables model extraction that determines the forms for a given data set. Three different testing approaches are used in information processing: guided and unrestricted study. [12] [12]. Classification is a first step in evaluating many situations. The overarching aim is to enhance market comprehension or forecasts. Analysis has proposed many common ways of classification techniques. Data mining 's central operation is designing accurate classification systems. A vital mission for data processing and machine learning research. Like decision-making trees, naive- Bayesian systems, serial minB.B.F.um optimization (S.M.O.), I.B.K., B.F. Chain modes of grouping methods, etc.

4. FEATURE SELECTION

Increased use of computers from both directions contributes to comprehensive data processing. This data are large, systematically interconnected data to identify acceptable patterns, making data mining a crucial area for data processing, prediction, and other activities. It has joined a complex science area to address real- time theoretical problems. Big data mining is used in many areas where data analysis is needed. These data mining and creation techniques were commonly used at various levels, such as pattern recognition, etc., and collecting apps plays an important role in virtually every field.

The collection attempts to assess the smallest possible subset of characteristics. The architecture selects the foundation of original features by removing redundant and unused dimensionality features without losing information. Until data-mining activities are implemented, the pre-processing step is critical. Mining accuracy, measurement time, and test comprehension are improved. Three filtering strategies comprises of philtres, wrappers and embedded approaches.

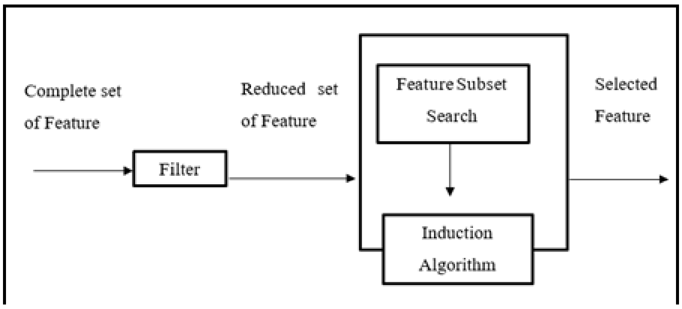


Fig. 1. Hybrid feature selection method [16]

As discussed in [16], the Filter selects the function without the classifier type used. The advantage of this method is that it is only necessary to select features once, it is simple and irrespective of the classifier used. This procedure, however, lacks the classifier relationship; each feature is interpreted separately from functional dependence. Wrapper's approach depends on classification. Classifier results are used to assess the goodness of the specified feature or attribute. Another method has the benefit that the filtering cycle eliminates the downside, which is simpler than the filtering system as it still takes all the dependencies. The next embedded approach is to combine a philtre algorithm with a wrapper approach to find an optimal sub-set in the classification structure. This method 's advantage is less expensive and less vulnerable than wrapper approach. Different feature selection applications over the past two decades:

- Text mining
- Image processing and computer vision
- Industrial applications
- Bioinformatics

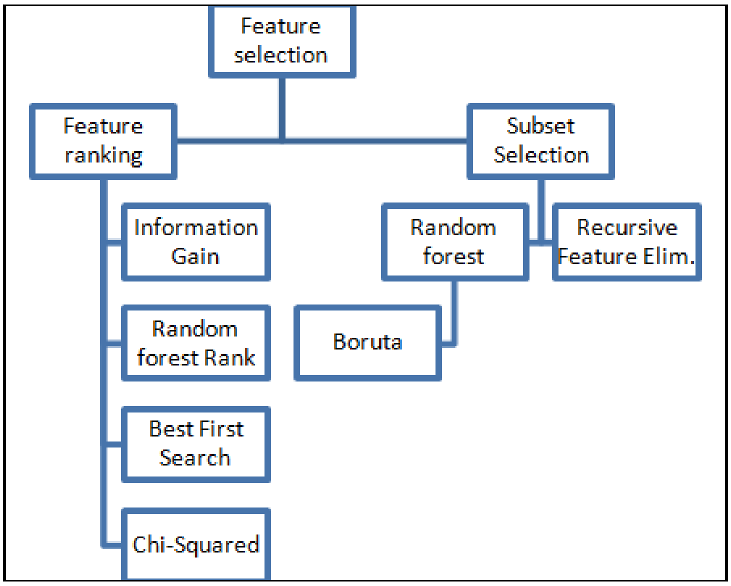


Fig. 2. Various Feature selection methods

5. Material and Methods

Weka is a data mining application that uses algorithms. Such algorithms can be directly applied to the data or code-named Java.

"Weka's a platform for:

- Regression.
- Clusters.
- Association.
- Pre-processing data.
- Classification Rating.
- Visualization."[13]

We have Weka classifiers to approximate measurable or numerical quantities. Decision and lists are available for learning systems, vector-support machines, case-dependent classifiers, technical regression, and Bayes networks. If loaded, all tabs are allowed. The best algorithm for basic data representation can be found for parameters, tests and errors.

The tab Cluster aims to identify clusters or event groups in data set. Clustering provides the customer details for analysis. Clustering uses the training set, percentage division, test set, and classes issued, for which users can ignore other requirement-based data set attributes. "K-Means, EM, Cobweb, X-means and Farthest First in Weka."

6. Proposed Method:

6.1. Sequential Minimal Optimization (S.M.O.)

The new teaching algorithm supporting vector machines (SVMs) is minimal sequential optimization. In 1998, John Platt developed it for minimal sequential optimization (S.M.O.)[14], a simple and easy approach for SVM preparation. The underlying theory is to solve the dual quadratic optimization by maximising the complete two-component sub-set.

S.M.O. 's advantage is the simple and clear introduction. A broad topic of quadratic programming is understanding a vector-supporting computer.

"S.M.O. distinguishes this core quadratic programming problem in a number of small potentials. Such a wide square

Analytically solved problem method, which avoids the use of time-consuming quadratic numerical programming as an internal loop. S.M.O. has a continuous memory feature, allowing S.M.O. to do intense exercises. Since matrix calculation is avoided, standard S.M.O. scales for different testing problems like linear and quadratic chunking SVM algorithm scales like linear and cubic. S.M.O. time is regulated by SVM evaluation; S.M.O. is also the fastest for linear, small data sets.

7. FRAMEWORK

Fig.3 provides an illustrative outline of SVM and gene-dependent clustering process. Below is a genetic clustering enhanced SVM process algorithm.

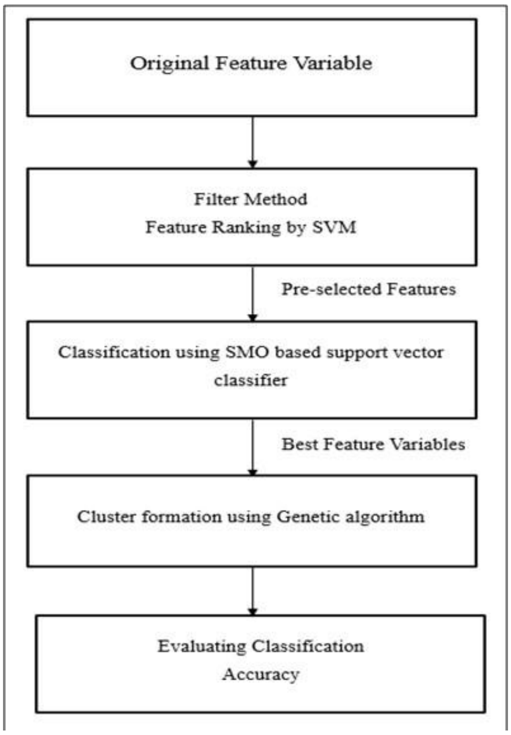
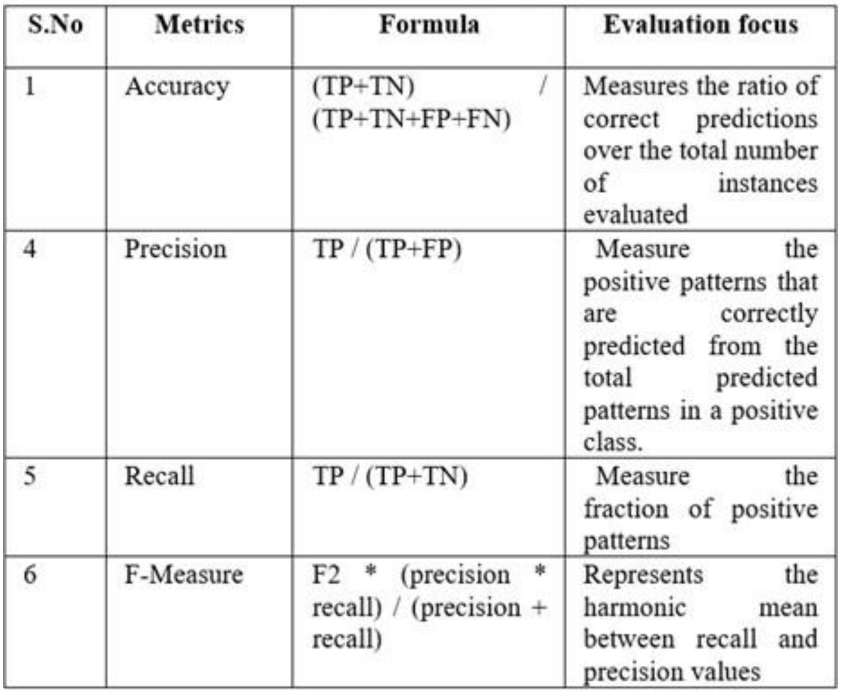

Original Feature Variable
Filter Method
Feature Ranking by SVM
Pre-selected Features
Classification using SMO based support vector classifier
Best Feature Variables
Cluster formation using Genetic algorithm
Evaluating Classification Accuracy

Fig. 3. Proposed Framework of S.M.O. classification with genetic algorithm
Select initial variables indexed by SVM philtre. Output dependent on adaptation, precision, estimation, memory, accuracy. In the data mining industry, these metric metrics played a crucial role in assessing the effects of different classifications and directing the algorithms as shown in Table I

Classification efficiency is calculated in the tables below.

Table I. Evaluation Parameters



S.No	Metrics	Formula	Evaluation focus
1	Accuracy	(TP+TN) / (TP+TN+FP+FN)	Measures the ratio of correct predictions over the total number of instances evaluated
4	Precision	TP / (TP+FP)	Measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.
5	Recall	TP / (TP+TN)	Measure the fraction of positive patterns
6	F-Measure	F2 * (precision * recall) / (precision + recall)	Represents the harmonic mean between recall and precision values

7.1. The Proposed Hybrid Feature Selection Algorithm

Within this section, as our observations indicate, we identify a primary genetic clustering S.M.O. algorithm. Several operations must be determined before using genetic algorithm. It is the first population, fitness, selection, and crossover.

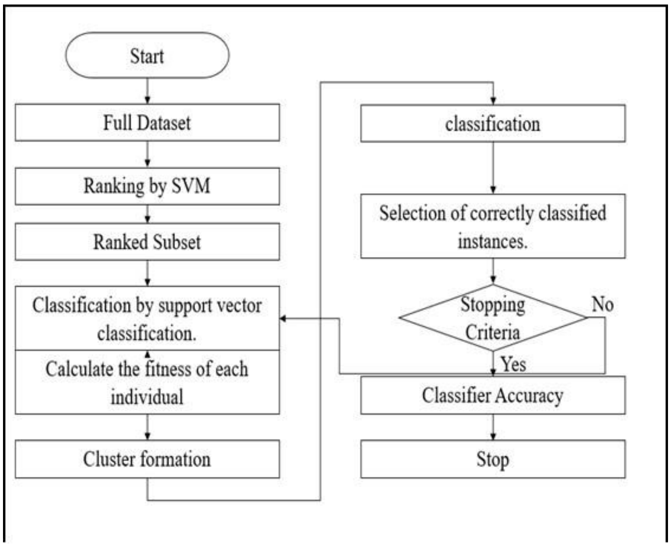

Start
Full Dataset
Ranking by SVM
Ranked Subset
Classification by support vector classification.
Calculate the fitness of each individual
Cluster formation
classification
Selection of correctly classified instances.
Stopping Criteria
No
Yes
Classifier Accuracy
Stop

Fig. 4. Proposed algorithm methodology

STEP 1: Population initialization: randomly generated initial population. A random function (g) generates a random number in [0 , 1] for each element. If the number exceeds a threshold, g=1. Otherwise, g=0. For example, the threshold value may be 0.5 to equate 1 or 0.

STEP 2: Minimal sequential vector recognition interface aid preparation.

STEP 3: It extracts all missing properties globally and converts them into binary attributes. This functions are also simply standardised. (This is important in the study of output parameters based on structured data, not original data.

STEP 4: The genetic algorithm integrates genetic algorithm K-means into a modern statistical clustering method.

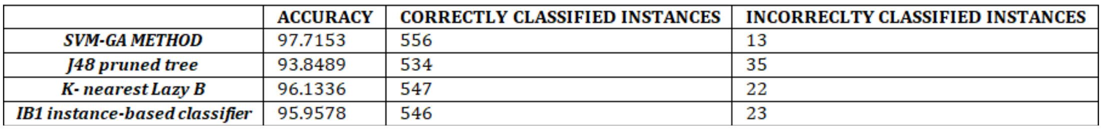
8. Results and Discussion

In our opinion, genetic algorithms are original methods that can be used for the extraction of functions. They suggest a system for the evaluation of individuals based on the combination of SVM[15] classifiers that have been qualified for each function. More specifically, we associate an SVM classifier with each primitive and implement a selection of classifiers. In the method we propose, classifier learning is Performed one step before each primitive genetic algorithm. The genetic algorithm fitness function is calculated by a mixture of these classes. We significantly reduce the training period for the that classifier. Therefore, the proposed selection approach significantly reduces implementation time. This dataset includes569 reports of breast cancer cases, 357 of whom were healthy, and 212 of whom were malignant. Class name and 32 function number are identical to the brain or malignant type of breast cancer. This functions are Measured by visual representation of the breast mass aspirates needle (F.N.A.) representing the cell nucleus characteristics in the photo.

8.1. S.M.O. Based SVM Algorithm Results

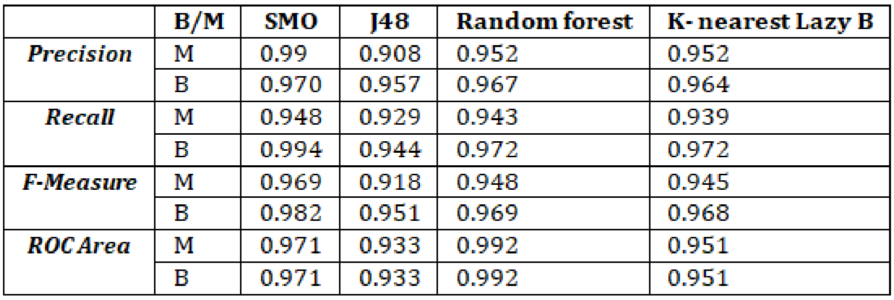
We run 5 K folded philtre output S.M.O. Classifier. Various other algorithms were contrasted with our system and can be seen below:

Table II. Result On Accuracy with Correctly and Incorrectly Classified Instances.



	ACCURACY	CORRECTLY CLASSIFIED INSTANCES	INCORRECLTY CLASSIFIED INSTANCES
SVM-GA METHOD	97.7153	556	13
J48 pruned tree	93.8489	534	35
K- nearest Lazy B	96.1336	547	22
IB1 instance-based classifier	95.9578	546	23

Table III. Comparison Of Various Algorithms On Precision, Recall, F-Measure, And Roc Area



	B/M	SMO	J48	Random forest	K- nearest Lazy B
Precision	M	0.99	0.908	0.952	0.952
	B	0.970	0.957	0.967	0.964
Recall	M	0.948	0.929	0.943	0.939
	B	0.994	0.944	0.972	0.972
F-Measure	M	0.969	0.918	0.948	0.945
	B	0.982	0.951	0.969	0.968
ROC Area	M	0.971	0.933	0.992	0.951
	B	0.971	0.933	0.992	0.951

Where, B = Benign, M = Malignant.
It can be shown that S.M.O. algorithm worked better in terms of Accuracy, Recall, F-measure, and R.O.C. field than other algorithms.

Table IV. Confusion Matrix Produced For Smo Classifier

a	b	
201	11	a=M
2	355	b=B

8.2. Genetic Clustering Algorithm Results

Founded by [15]. Standard Manhattan Distance class is used for outcome similarity calculations. Easy K-Means incorporates the basic missing qualities. If a mechanism creates a chromosome with all the records of one cluster, chromosomes are changed until at least two

Time is taken to build Genetic clustering model (full training data)	38.81 seconds

Here, the algorithm took 38.81 seconds to run.

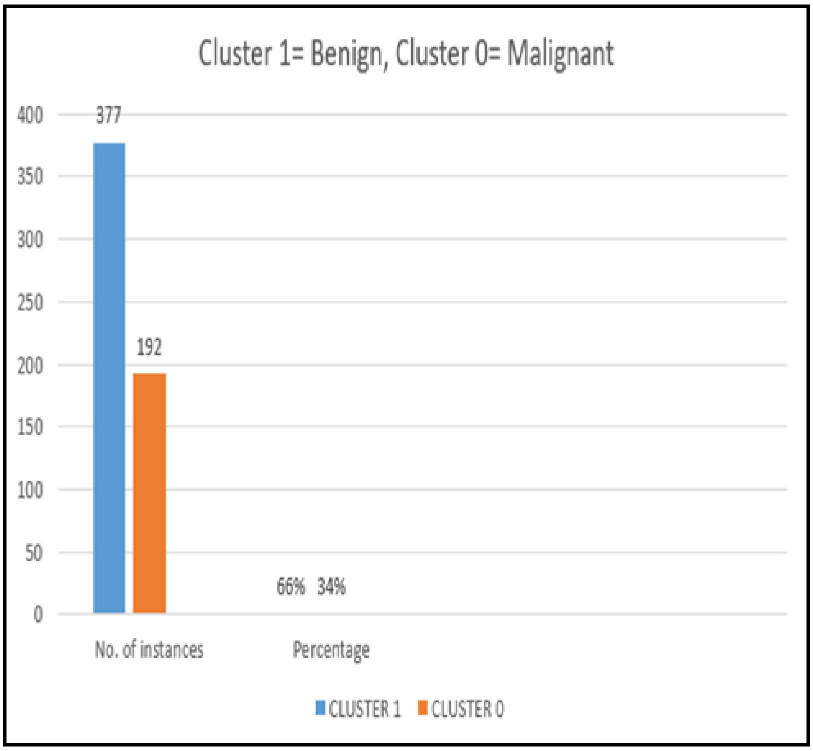
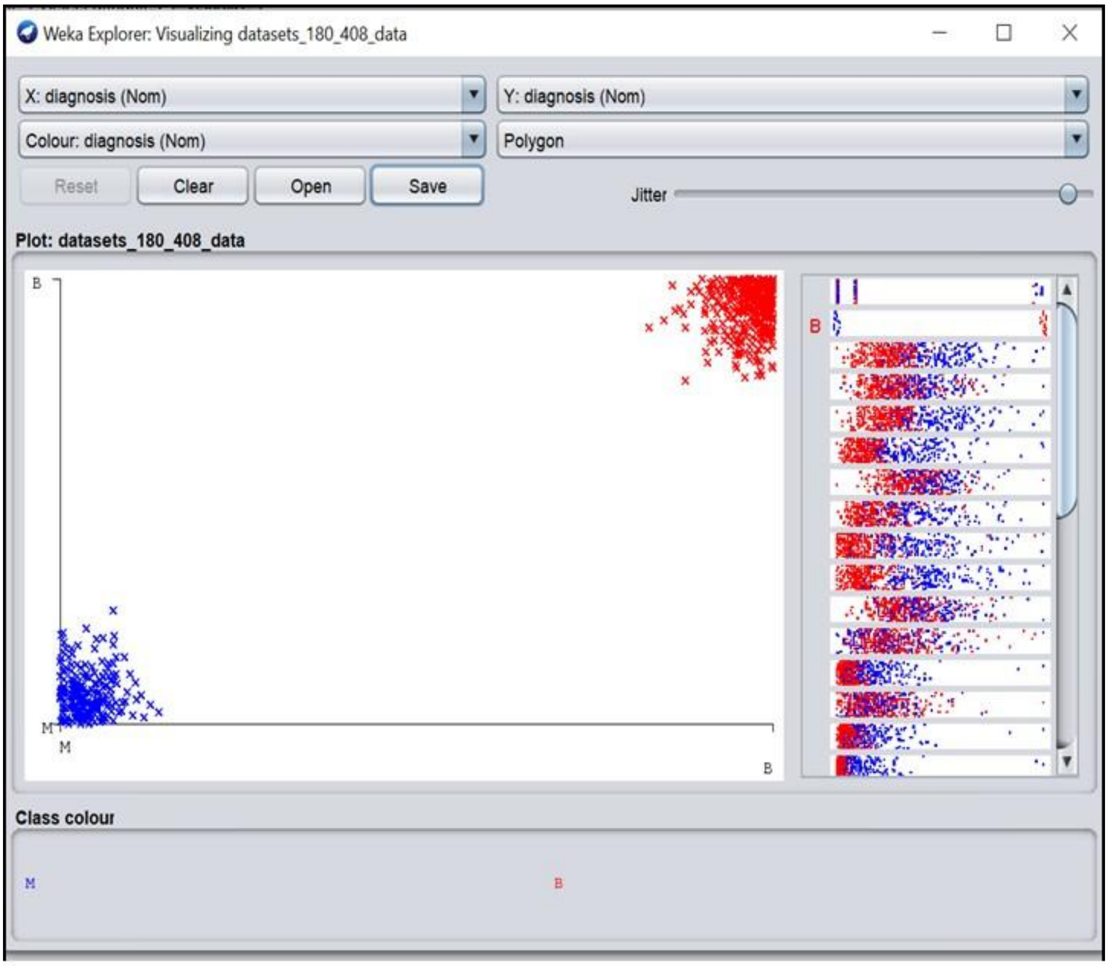


Fig. 5. Number of instances based on cluster formation

Table V. Confusion matrix of cluster

0	1	
184	28	M
8	349	B

Ultimately, it was found that the overall wrongly clustered instances is 36, 6.32 percent of the entire data collection. Visualization of cluster forming can be seen in the figure below indicating cluster formation selection.



Weka Explorer: Visualizing datasets_180_408_data

X: diagnosis (Nom)

Y: diagnosis (Nom)

Colour: diagnosis (Nom)

Polygon

Plot: datasets_180_408_data

Class colour

Fig. 6. Result on Genetic clustering model

Blue in the database shows malignant examples, and red in the Map indicates Benign cases. The plot reveals that clustering was successful.

9. Conclusion

This essay presents a new approach to feature collection. This selection approach is based on genetic algorithms combined with 5k fold-cross-validation SVM classification. Our system was evaluated on the Wisconsin dataset for breast cancer with several other forms classifier. We obtained very high accuracy results in grouping and quick results in clustering Breast-cancer diagnosis. This method not only provides a more accurate model estimation utility estimation, but also helps to assess the best ML algorithms parameters. Results show the suggested method's

robustness. Our goal is to launch the G.A. immediately. Analysis to ensure the required balance of consistency and time of diverse datasets from other fields of inquiry.

References

1. E.C. E.C. Fear, P.M. Meaney, and M.A.Stuchly,"Breast Cancer Detection Microwaves, "IEEE Potentials, vol.22, pp.12-18,February-March 2003.

2. Homer MJ.Mammographic: A realistic solution. Boston, MA, second edition, 1997.

3. Reston VA, Illustrated Breast Imaging Reporting and Data System (BI-RADSTM), third edition, 1998.

4. Urmaliya, Ajay, Singhai. (2013). (2013). Minimal sequential optimization to help vector machine for breast cancer diagnosis function collection. 2013 IEEE 2nd International Image Processing Conference, IEEE ICIIP 2013. 481–486. ICIIP.2013. 6707638 10.1109/2013.

5. Bittern, R., Dolgobrodov, D., Marshall, R., Moore, P., Steele, R. Artificial neural cancer networks. All Hands Conference 19 (2007), pp.251 – 263.

6. Bellaachia, A., AND Guven, E. Estimating breast cancer longevity through data mining.

7. Afef Ben Brahim, Mohamed Limam, 2017, "Ensemble feature selection for high-dimensional data: a new approach and comparative analysis," IFCS, vol. 12(4), 937-952

8. Huerta, E.B., 2006, "A Hybrid GA / SVM method for gene discovery and microarray data classification," springer, vol.3907 pp.34–44

9. Sallehuddin, R.N.H.aidillah, S.H., Mustaffa, N.H., 2014, "Liver cancer detection using artificial neural networks and supporting vector machines" In: Proceedings of the International Conference on Progress in Communication Network, and Computation, Elsevier Research, pp 487-493.

10. Jabbar, M.A., Deekshatulu, B.L., Chandra, 2012, "Heart disease prediction method using associative and genetic algorithm."

11. S. S. Nikam, "A Comparative Study of Data Mining Algorithms Classification Strategies," Oriental Computer Science & Technology Journal, vol.8, pp.13-19, April 2015.

12. S. Neelamegam, E.Ramaraj, "International Journal of P2P Network Developments and Technology (IJPTT), vol. 4, pp. 369-374, September 2013.

13. Eibe Frank, Mark A. Hall, Witten (2016). Workbench WEKA. "Data Mining: Functional Machine Learning Methods and Techniques," Morgan Kaufmann, Fourth Edition, 2016.

14. Fourteen. "Platt, J.C.: Minimum sequential optimization: quick algorithm for training vector machines. MSR-TR-98- 14, Microsoft Research, 1998.

15. Fifteen. 'Islam, M. Z., Estivill-Castro, V., Rahman, M. A. (From 2018). Combining K-Means and a Genetic Algorithm into a Novel Method for High Quality Clustering. Application-expert systems. 91:402-417 AM.

16. Sangaiah, Vincent Antony Kumar, A. (2019). (2019). Improving the efficiency of medical diagnosis using hybrid function selection by reliefs and entropy-based genetic search (RF-EGA) approach: breast cancer prediction. Computing array. 22. 22. 10.1007 / s10586-1702-5.