ISSN 2278-3091
Volume 8, No.1.6, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse1881.62019.pdf
https://doi.org/10.30534/ijatcse/2019/1881.62019

Ensamble Based Multi Filters Algorithm for Tumor Classification in High Dimensional Microarray Dataset

Tengku Mazlin Tengku Ab Hamid[1], Roselina Sallehuddin[2], Zuriahati Mohd Yunos[3], Aida Ali[4]
[1,2,3,4]School of Computing, Faculty of Engineering, University Teknologu Malaysia (UTM), Skudai, Johor, Malaysia

ABSTRACT

Prevalent adoption of machine learning has magnified its requirements in high dimensional microarray data classification. Due to explosive increase of data dimensionality, the existence of features redundancy and ambiguity directly leads to classification inaccuracy. Filter feature selection algorithms are capable to boost classification accuracy and diminish computational complexity by extracting relevant information through supervised learning. However, the independent filter algorithm is incompetent to consider the features interaction which resulting an imbalance selection of significant features and consequently degrading the classifier performance. This paper presents an assemblage of multi filters algorithm which assembles four filters algorithm outputs with frequency of occurrence rate evaluation to improve classification performance by attaining an optimal number of significant features. Experimental analysis was performed on a standard Breast Cancer dataset consists of 286 instances and Support Vector Machine (SVM) classifier. The experimental results proved that the ensemble based multi filters algorithm with occurrence rate evaluation successfully depletes from 9 original dataset features to 5 optimal significant features. The finding indicates that this technique competently signifies SVM classification performance in terms of accuracy with optimum significant features for high dimensional microarray data compared to independent filter algorithm.

Key words : High deminsional dataset, Feature selection, Ensamble based multi filters, algorithm, Classification, Support vector machine

1. INTRODUCTION

Classification of bioinformatics data is a significant machine learning tasks that have been widely adapted for high dimensional microarray dataset. For enhancing the classification accuracy, it is essential to identify features with highest importance as the classification performance is highly rely on the quality of the training data [1]. Due to the escalating amount of information needs to be processed, there exists some redundancy and unimportant features which consequently resulting in immoderate training and classification time [2].

Feature selection is a common approach implemented to address such explosive increase of features in high dimensional data issues and reducing information to improve the classification performance. Feature selection approach can be categorized into filter, wrapper and embedded techniques [3]. In filter techniques, attributes are grouped according to the inherent information without the use of classification algorithms [4]. Features are evaluated and ranked by each intrinsic quality using ranking calculations such as weights, dependency and distance measure. Such ranking techniques are certainly preferable for handling large dimensional datasets [4]. However, the single filter technique is still incompetent in considering the interaction between features which resulting an unbalanced selection of significant features [5]. In contrast, even though the wrapper and embedded techniques may produce more precise result, both techniques are much time consuming as certain classification algorithm is required in evaluating the sets of features [6]. Thus, the need for intelligent feature selection technique is required to examine features with highest significance and reliable to precise the classification tasks as well as eliminating the redundant features efficiently.

Therefore, this paper presents an assemblage based multi filters feature selection techniques that assembles four filters algorithm utilization outputs using Information Gain (IG), Gain Ratio (GR), Chi-squared (CS) and Relief-F (RF) with features occurrence rate evaluation to significantly improve the classification accuracy while predicting the optimal number of significant features reliable for the classification process of SVM classifier.

The structure of this paper is sorted as follows: Section 2 presents the related research on feature selection techniques. Section 3 described all processes and techniques involved in the research methodology. Next, the experimental results and findings are discussed in Section 4. Finally, Section 5 concludes the paper.

2. RELATED RESEARCH ON FEATURE SELECTION TECHNIQUES

Significant features eminently influence the classification performance of biomedical data due to the importance of the medical information. According to

reviews on feature selection, the performance of classifier is mostly relied on the quality of the training data which used to train the classifier such as SVM [6,7]. For more robust and accurate classification, feature selection techniques have been widely utilized to encounter the data dimensionality issues where the information is extracted from the original large data to a reduced set of significant features. In addition, features quality and model selection are the two main components that need to be considered in developing an optimal classification model [8]. However, selecting optimum significant features from high dimensional data may produce a challenge especially to an overfitted data that consequently resulting to data dimensionality issues [6,7,8].

In recent years, many feature selection techniques have been conducted on filter, wrapper and embedded approach in attempts to eliminate the irrelevant redundancy issues in high dimensional microarray datasets. A study in [9] proposed a filter based technique using the Maximal Information Coefficient (MIC) and Gram-Schmidt Orthogonalization (GSO) or named as Orthogonal MIC Feature Selection (OMICFS). In this approach, the relevance between feature variables is quantified using MIC and the GSO is used to evaluate the orthogonalized variable of feature with respect to previous selected features. The results of orthogonalization strategy allows OMICFS to eliminate the irrelevant redundancy without any additional process. The work of [10] proposed a filter based technique using jack knife iteration and voting classifier for selecting significant features. Features are ranked based on the absolute value of t-statistic calculated with the remaining training sample and selected feature with the highest t-statistic is constructed into the voting classifier. The result shows an increased in accuracy performance when 1% percent of top ranked features were used in the classifier compared to 5% percent of top ranked features.

A report from [11] proposed a multi heuristic based filter techniques using X-variance and Mutual Congestion for gene selection problems in binary medical data. X-variance is evaluated based on the subset of features such as mean and variance whereas Mutual Congestion is calculated based on the feature's frequency. The results show that the accuracy of the independent classifiers in high dimensional data significantly improved using Mutual Congestion compared to X-variance which achieved comparable result. This indicates that the techniques which classify by subset of features are more reliable for low dimensional data while the frequency based are more reliable for high dimensional data. Mean while, the work of [12] proposed a multi filter techniques using the concepts of Mutual Information (MI), RF and Fisher Score (F-Score) to solve the redundancy problems. RF and F-Score are used to determine the highest ranked features while providing the mutual relevance between features instead of using the mutual redundancy. For feature selection, MI technique using maximum relevance and minimum redundancy is adopted in single and multi objective differential evolution algorithms. The results outperform MI in both single and multi objective differential evolution frameworks which indicates that the classification performance can be improved if feature selection is considered as a multi objective problem in terms of the size of feature subset and the classification accuracy. A reported work in [13] also proposed a multi filter techniques by combining several statistical measures to determine the groups of potential biomarkers for lung cancer. The result shows better discriminative ability and convenient for genes selection.

A study from [14] proposed a filter and wrapper based techniques which a new distance based evaluation function is utilized if the same class samples are attracted to each another, whereas different class samples are far apart, and a set of candidate feature subsets is determined using a weighted bootstrapping search strategy. In order to select the optimum features, specific classifier and cross validation were used to validate the performance. The results proved that the overfitting between the optimal feature subset and a given classifier have significantly overcomed. Other combination of filter and wrapper techniques is proposed in [15] to measure the significance of feature using a majority voting where ten different filter and wrapper algorithms were utilized. The results increased the accuracy performance compared to the single filter feature selection even though the influence of wrapper techniques has increased some computational complexity as the wrapper techniques acquire specific learning algorithms to perform the selection process [15].

Based on the mentioned related works, suggestions of assembling feature selection techniques has significantly improved the accuracy performance of classification algorithms. Filter techniques is highly recommended due to its efficiency of computationally fastest in handling large datasets with the use of ranking and space searching technique. Furthermore, features that is independently weak but strong as a group can be identified, the redundant features can be eliminated and the highest correlated features with the output class can be determined by utilizing the hybridization of feature selection techniques [6,14,15].

Thus, this research is motivated to present an ensemble based multi filters feature selection techniques which assembles four filters algorithm utilization outputs using IG, GR, CS and RF with frequency rate of occurrence evaluation to determine the most optimum significant features. For classification process, SVM classifier is adopted due to its robust accuracy performance in classifying high dimensional data as suggested in reported studies. An overall classification performance is assessed using a standard Breast Cancer dataset and evaluation metrics in terms of accuracy, specificity, sensitivity and AUC.

3. RESEARCH METHODOLOGY

Selecting features using filter algorithms are usually processed by assigning a score or rank to each input features and the ranked features are then can be selected into

the classifier or eliminated according to the scored value. It is executed independently without including other classifier algorithm, which made it computationally efficient due to time complexity reduction and capable in handling various datasets dimension. In this section, all techniques that involved in the process of assemblage of multi filters feature selection algorithm with frequency of occurrence rate evaluation are described in details.

3.1. Information Gain Utilization

Information Gain (IG) is utilized by minimizing the unreliability of features corresponded with finding the unspecified value of a class attribute [16]. It is measured using a distribution of entropy value according to the significance in observing different class. Features are ranked correspond to the evaluated entropy value regarding to its class. The evaluation of entropy value is derived using Equation 1:

$$IG\ (x_i \mid y_i) = H(x_i) - H(x_i \mid y_i) \qquad (1)$$

Where, x defined a class given information y. A feature is considered as significant if the IG value is smaller than the entropy value and vice versa [16].

3.2. Gain Ratio Utilization

Gain Ratio (GR) provides bias improvement of IG especially for features with larger diversity value [17]. It overcome bias by considering important information and uses branches size to identify features. GR indicates a small value when all information belongs to single branch while a bigger value when the information is spread evenly [17]. The features will be splitted into a ratio and the significance features is selected using Equation 2:

$$GR\ (x_i) = Gain\ (x_i) \div SplitInfo\ (S, x_i) \qquad (2)$$

Features with the highest GR value are selected as the splitting features and considered as the significant information [17].

3.3. Chi-squared Utilization

Chi-squared (CS) evaluates the features significance by measuring the dependency between features with respect to its class using a statistical calculation. A CS score is calculated to test the dependency level between features based on their observed and expected value [18]. CS evaluation is derived using Equation 3:

$$x^2 = \sum (observed_i - expected_i)^2 \div expected_i \qquad (3)$$

Features with highest significance level indicates a high feature dependency and significant [18]. Otherwise, it is indicated as independent features with less correlation and thus can be eliminated from the learning tasks.

3.4. Relief-F Utilization

Relief-F (RF) is capable for handling ambiguity and incomplete data as well as multi class problem [19]. RF outperformed other filter techniques due to its low bias and suitable to be applied in any domains. The ranking of features is done by measuring the difference value or weighted score between features towards their nearest neighbour as derived in Equation 4:

$$W_i = W_i - (x_i - nearHit_i)^2 + (x_i - nearMiss_i)^2 \qquad (4)$$

Features in the similar class is recognized as a 'hit', with a low difference value. Alternatively, the features with high difference value are in the opposite class and recognized as a 'miss' [19]. A low difference value is considered as significant feature in which the strength between neighbours are attracted towards each other. Whereas, a less significant feature is represented by a large difference value with low strength of neighbouring instances and less attraction to each other.

3.5. Ensemble Based Multi Filters Algorithm Outputs Utilization Process

The assemblage of multi filter algorithms process is done by utilizing IG, GR, CS and RF algorithms to obtain four sets of ranked features. In order to select the ranked features, the threshold value used are the entropy value, gain ratio value, significance level and feature score which are set to 0.05 [16,17,18,19]. Next, these utilization outputs are assembled, and a simple evaluation on features occurrence rate is computed to determine the most optimal significant features prior to the learning tasks. As for optimum features selection, the maximum frequency of occurrence rate is set to 4 [9,12,13]. A counter is used to perform the selection of assembled features according to the computed occurrence rate. Figure 1 illustrates the assemblage of multi filter algorithms utilization process and each process are explained in detail accordingly.

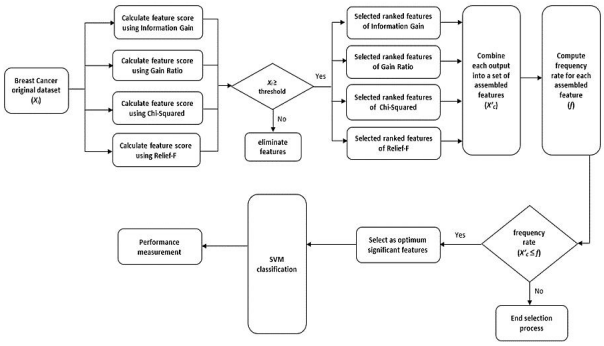

Figure 1: Process of Ensemble Based Multi Filter Algorithms Feature Selection

3.5.1 Algorithm 1: Multi Filter Algorithms Utilization Ranking

Initially, four filter techniques are utilized using IG, GR, CS and RF algorithms on the original dataset to evaluate the ranking score of each attribute. A threshold value is used to select the sets of ranked features. The algorithm to utilize the four filter algorithms ranking is described as follows:

- Step 1: Derive the original Breast Cancer dataset as $X_i = \{X_1, X_2, \ldots, X_{286}\}$ and $C_i = \{C_1, C_2\}$ as the class of benign and malignant.
- Step 2: Evaluate X_i ranks using IG, GR, CS and RF algorithms according to each ranking evaluation.
- Step 3: Set 0.05 as the threshold value to select the sets of ranked features in Step 2.
- Step 4: Select X_i based on the threshold value and produce four sets of selected ranked features as X'_i.

3.5.2 Algorithm 2: Assemble of Multi Filter Algorithms Utilization Outputs and Occurrence Rate

From Algorithm 1, the four utilization outputs are combined to produce a set of assembled features. Then, the occurrence rate for each assembled feature is computed. The algorithm to assemble the four utilization outputs is described as follows:

- Step 1: Obtain four sets of X'_i outputs from Algorithm 1.
- Step 2: Combine each X'_i into a set of assembled features as X'_C.
- Step 3: For each X'_C, compute the occurrence rate (*f*).

3.5.3 Algorithm 3: Assemble Selection for Optimum Significant Features

From Algorithm 2, the occurrence rate for each assembled feature is evaluated to select the most optimum significant features. The algorithm to evaluate the occurrence rate of assembled features is presented as follows:

- Step 1: Set 4 as the maximum frequency of occurrence rate.
- Step 2: Test X'_C with each occurrence rate and find the intercepts.
- Step 3: Select X'_C as the optimum significant features until the maximum is achieved.
- Step 4: Perform classification using SVM and validate the accuracy performance.

3.6. Support Vector Machine (SVM) Classification

Support Vector Machine (SVM) is adopted as a supervised learning classifier which maximize the margin between data in a set of samples of two classes [20]. The input data are divided by an optimal hyperplane and kernels are used to transfer the input data to a high dimensional space where a hyperplane partition can be established [21]. The objective of the learning algorithm is to determine the SVM parameters until where the space dimension is maximized [22-24]. Since an overfitting SVM may easily affected when the error is too high or too low, an optimal partitioning hyperplane is established as the optimal decision surface which formed using Equation 5 [23]:

$$f(x) = sgn\left(\sum L_i\left(y_i\left(x_i, x\right)\right) + b_0\right) \qquad (5)$$

Where, *sgn* defined a symbolic function, L_i represents an optimal Lagrange coefficient and the optimum value of SVM parameters are denoted by (x_i , x) and b_0. According to reported studies, Radial Basis Function (RBF) is the most frequently used kernel function due to its localization ability and finite response along the hyperplane. RBF kernel function can be derived as in Equation 6 [24]:

$$k(x, x') = exp\left(-\|x - x'\|^2 \div 2\sigma^2\right) \qquad (6)$$

Where, σ defined the kernel width and (x, x') defined the kernel parameter. For the training process in this research, the SVM classification algorithm using RBF kernel function with 10-fold cross validation is adopted to measure the overall classification analysis performance.

3.7. Experimental Dataset

A standard Breast Cancer dataset consisting of 286 instances with 9 attributes of features and 2 classes which defined as benign or malignant is used for the experiment and analysis. This dataset is available and can be obtained from the UCI Machine Learning Repository [25]. Each attribute is notated according to each feature's name for understandable analysis. Table 2 shows the details summary of Breast Cancer dataset features.

Table 1: Details summary of Breast Cancer Dataset Features

Attributes	Features Notation	Features Name	Details
1	A1	Age	Patient's age at the time of diagnosis.
2	A2	Menopause	Patient's menopausal state (pre or postmenopausal).
3	A3	Tumour Size	Maximum diameter of the excised tumor (in mm).
4	A4	Inv-Nodes	Number of axillary lymph nodes with metastatic state (range 0-39).
5	A5	Node Caps	Capsule of the lymph nodes with metastatic state.
6	A6	Degree of Malignancy	Histological grade of the tumor (range 1-3).
7	A7	Breast	Side location of the breast (right or left).
8	A8	Breast Quadrant	Quadrant of the breast.
9	A9	Irradiation	Radiation therapy treatment.

3.8. Performance Measurement

SVM classification performance are evaluated using evaluation metrics in terms of accuracy, sensitivity, specificity, and Area under Receiver Operating Characteristic Curve (AUC). Based on statistical equations, measures of true positive (TP) refers to the number of correctly classified malignant data while true negative (TN) refers to the percentage of correctly classified benign data. False positive (FP) represents a data which classified as malignant when it is benign, while false negative (FN) represents the misclassification of data which classified as benign when it is malignant [15,21,22,23]. Table 2 describes the evaluation metrics such as accuracy, sensitivity, specificity, and AUC used to validate the classification performance of SVM classifier.

Table 2: Evaluation Metrics used in Performance Measurement

Evaluation Metrics	Statistical Equations
Accuracy	(TP + TN) ÷ (TP + FN + TN + FP) × 100
Sensitivity	TP ÷ (FN + TP) × 100
Specificity	TN ÷ (FP + TN) × 100
AUC	0.5 (TP ÷ (FN + TP)) + (TN ÷ (FP + TN)) × 100

4. RESULTS AND ANALYSIS

This section discussed on the results obtained from the ensemble based multi filter algorithms utilization process in determining the optimum significant features. The analysis was carried out on a standard Breast Cancer dataset to validate the classification performance of the assemblage process using the SVM classifier.

4.1. Multi Filter Algorithms Utilization Ranking Outputs

In the first phase, multi filter feature selection was used to produce four sets of ranked features from the original dataset by utilizing four filter algorithms such as IG, GR, CS and RF. Features are ranked based on the different ranking evaluation and selected as significant features according to the threshold value used in each filter algorithm. Based on the four filter algorithms utilization, 9 features from the original dataset are ranked using IG, GR, CS and RF algorithms. A feature selection algorithm is claimed to verify the dataset if each feature obtained a high ranking score. For each feature, the ranks of each filter algorithm is calculated using specific ranking evaluation to produce the average ranking score. Features that complement the threshold value are selected as the significant features. Table 3 shows the average ranking score and the selected ranked features obtained from the IG, GR, CS and RF algorithms utilization.

Table 3: Multi Filter Algorithms Ranking Score and Selected Outputs

Ranks	Multi Filter Algorithms Ranking				Average Ranking Score			
	IG	GR	CS	RF	IG	GR	CS	RF
1	**A6**	**A5**	**A6**	**A6**	**0.078**	**0.071**	**28.875**	**0.093**
2	**A4**	**A4**	**A4**	**A8**	**0.071**	**0.054**	**26.594**	**0.062**
3	**A3**	**A6**	**A5**	**A2**	**0.061**	**0.051**	**19.731**	**0.057**
4	**A5**	A9	**A3**	**A1**	**0.051**	0.033	**17.039**	**0.051**
5	A9	A3	**A9**	A7	0.026	0.02	**9.792**	0.048
6	A1	A1	**A1**	A3	0.012	0.006	**3.956**	0.05
7	A8	A8	A8	A9	0.01	0.005	3.462	0.033
8	A7	A7	A7	A5	0.003	0.002	0.887	0.026
9	A2	A2	A2	A4	0.003	0.003	0.94	0.018

Based on Table 3, note that features from the first rank order are considered as highest significance and vice versa. The multi filters algorithm utilization process observed that several selected features are similar even though each filter algorithm perform different ranking evaluation. By setting 0.05 as the threshold value, IG and RF algorithm selected 4 features, GR algorithm selected 3 features while CS algorithm selected 6 features out of 9 original features. Feature A7 is not selected by the four filter algorithms since it obtained less than the threshold value, which can be eliminated. These four filters algorithm outputs indicate an unbalance number of selected features to be considered as significant. Thus, the occurrence rate evaluation is needed to optimize the selection of significant features.

4.2. Assembled Features Output and Occurrence Rate Evaluation

Four filters algorithm utilization outputs are then combined into a set of assembled features according to the ranking score respectively. The maximum threshold for frequency of occurrence rate is set to 4 and the intercept for each assembled feature is identified. Table 4 shows the set of assembled features with each computed occurrence rate.

Table 4: Assembled Features Output and Occurrence Rate

Ranks	Selected Ranked Features				Assembled Features	Occurrence Rate (*f*)			
	IG	GR	CS	RF		1	2	3	4
1	A6	A5	A6	A6	**A6**	✓	✓	✓	✓
2	A4	A4	A4	A8	**A5**	✓	✓	✓	-
3	A3	A6	A5	A2	**A4**	✓	✓	✓	-
4	A5	A9	A3	A1	**A8**	✓	-	-	-
5	A9	A3	A9	A7	**A2**	✓	-	-	-
6	A1	A1	A1	A3	**A3**	✓	✓	-	-
7	A8	A8	A8	A9	**A1**	✓	✓	-	-
8	A7	A7	A7	A5	**A9**	✓	-	-	-

Based on the results in Table 4, a set of assembled features with evaluation of occurrence rate have successfully computed. It is observed that features A6, A5, A4, A8, A2, A3, A1 and A9 signify the level of significance in which these features are observed to be occurring not more than 4 frequency across the four filters algorithm and achieved the intercept level. As the frequency increase, less features are observed to occur in the algorithm. Thus, an assemble selection is required to determine the most optimum features significant to improve the training process.

4.3. Optimum Significant Features of Assemble Selection

In assemble selection process, the outputs of assembled features with occurrence rate are trained using SVM classifier to determine the optimum significant features and the results is shown in Table 5.

Table 5: Optimum Significant Features based on Assemble Selection

Occurrence Rate (*f*)	Selected Optimum Features	Total Dataset Features	Total Selected Features	SVM Accuracy (%)
1	A6, A5, A4, A8, A2, A3, A1, A9	9	8	70.63
2	**A6, A5, A4, A3, A1**	**9**	**5**	**72.91**
3	A6, A5, A4	9	3	69.58
4	A6	9	1	69.93

Features A6, A5, A4, A3 and A1 achieved the highest SVM accuracy with 72.91% in the occurrence rates of 2 compared to other features in different frequency rate which achieved lower accuracy. The five optimum significant features are observed as Degree of Malignancy, Node Caps, Inv-Nodes, Tumour Size, and Age. From the assemblage of IG, GR, CS and RF algorithm outputs utilization, these five features are selected and considered as the most optimum features and significant to be include in the training process.

4.4. SVM Classification Performance and Validation

SVM classifier is trained to validate the classification performance in terms of accuracy, sensitivity, specificity and AUC. Table 6 shows the overall classification performance achieved by SVM classifier using the original dataset with 9 features, four sets of independent filters algorithm and the set of assembled features with 5 optimum significant features. 10-fold cross validation is applied on the input dataset and overall performance measures are averaged.

Table 6: Overall SVM Classification Performance on Breast Cancer Dataset

Filter Algorithms	Total Selected Features	Occurrence Rate	Overall SVM Performance (%)			
			Accuracy	Sensitivity	Specificity	AUC
None	9	None	56.69	0.889	0.024	0.491
IG	4	None	66.20	0.960	0.106	0.533
GR	3	None	69.10	0.920	0.247	0.584
CS	6	None	67.60	0.990	0.350	0.513
RF	4	None	68.90	0.955	0.470	0.510
Assembled Set	**5**	**Yes**	**72.91**	**0.993**	**0.447**	**0.627**

The results of SVM classification performance demonstrate that the assembled set using four filter algorithms utilization signify the SVM accuracy level with 72.91% compared to the accuracies achieved by other independent filter algorithms which produce an unbalance number of selected features. In addition, the rate of classification performance using ensemble based multi filters algorithm with occurrence rate evaluation techniques increase slightly higher by selecting 5 optimum significant features instead of 9 original features. This indicates that SVM classifier with ensemble based multi filter algorithms utilization could competently classified data correctly than the independent filter algorithm.

Based on sensitivity, the assembled set achieved the highest value by 99.30%. The highest sensitivity is important as it represents the percentage of correctly classified malignant data which indicate a patient might possible to have cancer and appropriate treatment can be taken immediately. AUC is used to evaluate the total effectiveness of correctly classified TP (malignant) and TN (benign) classes. Results observed that the AUC value in assembled set increased with 62.70% compared to the other filter algorithm. This indicate the SVM classification using assembled sets of four filters algorithm utilization could accurately determine the class of tumours. Figure 2 illustrates the overall SVM classification performance achieved by each filter algorithm utilization with 10-fold cross validation.

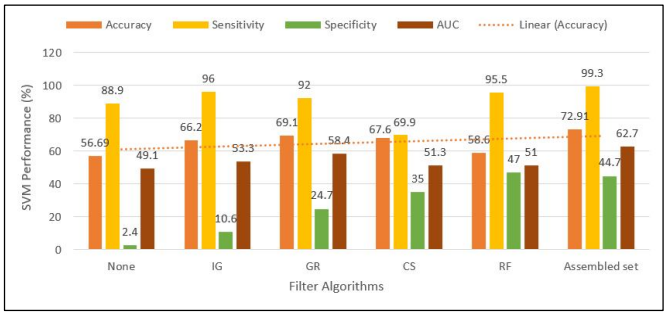

Figure 2: Overall SVM classification performance on Breast Cancer dataset

For overall effectiveness, the SVM classifier which trained with 5 optimum significant features could significantly improve the accuracy performance in classifying malignant and benign tumours. This suggests that the selection of optimum significant features with ensemble based multi filters algorithm utilization process is

highly essential and competently beneficial for establishing an intelligent and accurate SVM classifier for tumour diagnosis.

5. CONCLUSION

Requirement for intelligent classification of tumours has resulting a complexity in accuracy due to massive increment of high dimensional microarray data with the occurrence of redundant information. Feature selection techniques are proven as critical and convenient approach that need to be concerned when developing a classification system in which it is important for observing the optimum significant features to improve the performance of classification.

In this study, an ensemble based multi filters algorithm using IG, GR, CS and RF algorithm are utilized in determining the most optimum significant features on standard Breast Cancer dataset. Initially, features are ranked according to each filter algorithm's ranking score evaluation and threshold values to produce four sets of selected ranked features outputs. Then, these four outputs are assembled into a set of assembled features, and the optimum significant features are determined using a counter that evaluate the frequency of features occurrence rate across the four filters algorithm. Based on the assemble selection, 5 out of 9 features such as Degree of Malignancy, Node Caps, Inv-Nodes, Tumour Size, and Age were identified as the most optimum features and significant to be included in the SVM training process.

The results of overall SVM classification performance are validated by the prove that the ensemble based multi filters algorithm utilization process significantly improved the performance of SVM classification with optimum number of significant features by achieving 72.91% of accuracy, 99.30% of sensitivity, 44.70% of specificity and 62.70% of AUC. This work demonstrated that the assemblage of multi filters algorithm utilization technique using IG, GR, CS and RF and SVM classifier could achieve better accuracy scores with the selection of optimal significant features compared to the independent filter algorithm. Furthermore, the elimination of irrelevant and redundant features can be efficiently computed without degrading the classification accuracy as well as improving the knowledge discovery in high dimensional microarray data for better tumour diagnosis and treatments.

ACKNOWLEDGEMENTS

This research is supported by Research University Grant, Universiti Teknologi Malaysia (UTM) under vote number Q.J130000.2528.16H57 (GUP TIER 1:16H57), Q.J130000.2628.14J13 (GUP TIER 2:14J13) and RJ130000.7828.4F989. The authors would like to thank the anonymous reviewer for providing constructive and generous feedback. The authors would like to state deepest gratitude to Research Management Centre (RMC) of UTM, for the helpful research activities which indirectly improve the research process and Applied Industrial Analytics (ALIAS) research group for the support and motivation in making this research a success.

REFERENCES

1. R. R. Rani and D. Ramyachitra, "Microarray Cancer Gene Feature Selection using Spider Monkey Optimization Algorithm and Cancer Classification using Support Vector Machine," *2018 8th International Conference on Advances in Computing and Communication. ICACC 2018*, vol. 143, pp. 108-116, 2018. https://doi.org/10.1016/j.procs.2018.10.358
2. B. K. Singh, *et al.*, "Integrating Radiologist Feedback with Computer Aided Diagnostic Systems for Breast Cancer Risk Prediction in Ultrasonic Images: An Experimental Investigation in Machine Learning Paradigm," *Journal of Expert Systems with Applications*, vol. 90, pp. 209-223, December 2017. https://doi.org/10.1016/j.eswa.2017.08.020
3. H. Ghimatgar, *et al.*, "An Improved Feature Selection Algorithm Based on Graph Clustering and Ant Colony Optimization," *Journal of Knowledge-Based Systems*, vol. 159, pp. 270-285, November 2018. https://doi.org/10.1016/j.knosys.2018.06.025
4. P. R. Javier, *et al.*, "A General Framework for Boosting Feature Subset Selection Algorithms," *Journal of Information Fusion*, vol. 44, pp. 147-175, March 2018.
5. J. Wu, *et al.*, "Feature Selection for Cancer Classification Using Microarray Gene Expression Data," *Journal of Biostatistics and Biometrics*, vol. 1, pp. 1-7, April 2017.
6. J. Miao and L. Niu, "A Survey on Feature Selection," *Journal of Procedia Computer Science*, vol. 91, pp. 919-926, 2016. https://doi.org/10.1016/j.procs.2016.07.111
7. N. Bi, *et al.*, "High-Dimensional Supervised Feature Selection via Optimized Kernel Mutual Information," *Journal of Expert Systems with Applications*, vol. 108, pp. 81-95, October 2018.
8. J. Agor and O. Y. Ozaltin, "Feature Selection for Classification Models via Bilevel Optimization," *Journal of Computers & Operations Research*, vol. 106, pp. 156-168, June 2019. https://doi.org/10.1016/j.cor.2018.05.005
9. H. Lyu, *et al.*, "A Filter Feature Selection Method Based on the Maximal Information Coefficient and Gram-Schmidt Orthogonalization for Biomedical Data Mining," *Journal of Computers in Biology and Medicine*, vol. 89, pp. 264-274, August 2017.
10. S. L. Taylor and K. Kim, "A Jackknife and Voting Classifier Approach to Feature Selection and Classification," *Journal of Cancer Informatics*, vol. 10, pp, 133-147, 2011.

11. M. Alirezanejad, *et al.*, "Heuristic Filter Feature Selection Methods for Medical Datasets," *in Genomics*, 2019, pp. 1-16, doi: 10.1016/j.ygeno.2019.07.002.
12. E. Hancer, *et al.*, "Differential Evolution for Filter Feature Selection Based on Information Theory and Feature Ranking," *Journal of Knowledge-Based Systems*, vol. 140, pp. 103–119, November 2017.
13. S. Vítor, *et al.*, "Ensemble Feature Ranking Applied to Medical Data," *Journal of Procedia Technology*, vol. 17, pp. 223-230, 2014. https://doi.org/10.1016/j.protcy.2014.10.232
14. J. Zhang, *et al.*, "A New Hybrid Filter/Wrapper Algorithm for Feature Selection in Classification," *in Analytica Chimica Acta*, 2019, pp. 1-41, doi: 10.1016/j.aca.2019.06.054.
15. B. K. Singh, *et al.*, "Fuzzy Cluster Based Neural Network Classifier for Classifying Breast Tumors in Ultrasound Images, " *Journal of Expert Systems with Applications*, vol. 66, pp. 114-123, December 2016.
16. B. S. Pardo, *et al.*, "On Developing An Automatic Threshold Applied to Feature Selection Ensembles," *Journal of Information Fusion*, vol. 45, pp. 227-245, January 2019.
17. M. F. Tresna, *et al.*, "Data Mining Approach for Breast Cancer Patient Recovery," *EMITTER International Journal of Engineering Technology*, vol. 5, pp. 36-71, June 2017. https://doi.org/10.24003/emitter.v5i1.190
18. A. D. Seema, *et al.*, "A Feature Selection Approach for Enhancing the Cardiotocography Classification Performance," *Journal of Engineering and Techniques*, vol. 4, pp. 222-226, April 2018.
19. Z. M. Hira and D. F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data," *Journal of Advances in Bioinformatics*., pp. 1-13, July 2015.
20. J. U. Ryan, *et al.*, "Relief-Based Feature Selection: Introduction and Review," *Journal of Biomedical Informatics*, vol. 85, pp. 189-203, September 2018.
21. S. Roselina, *et al.*, "An Improvement in Support Vector Machine Classification Model using Grey Relational Analysis for Cancer Diagnosis," *Journal of Technology (Science & Engineering)*, vol. 78, pp. 107-119, March 2016.
22. S. Huang, *et al.*, "Applications of Support Vector Machine Learning in Cancer Genomics," *Journal of Cancer Genomics & Proteomics*, vol. 15, pp. 41-51, January 2018. https://doi.org/10.21873/cgp.20063
23. M. Xi, *et al.*, "Cancer Feature Selection and Classification using A Binary Quantum-Behaved Particle Swarm Optimization and Support Vector Machine," *Journal of Computational and Mathematical Methods in Medicine*, pp. 1-9, July 2016. https://doi.org/10.1155/2016/3572705
24. X. Wang, *et al.*, "Classification of Spot-Welded Joint Strength using Ultrasonic Signal Time Frequency Features and PSO-SVM Method," *Journal of Ultrasonics*, vol. 91, pp. 161-169, January 2019.
25. M. Zwitter and M. Soklic, Breast Cancer Dataset, Institute of Oncology University Medical Center Ljubljana, Yugoslavia, January 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer
26. Dhariwal, Sumit, and Sellappan Palaniappan. “An Efficient Approach for Semantic Image Classification Using Normalization Method.” International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no. 4, 2019, pp. 1268–74 https://doi.org/10.30534/ijatcse/2019/37842019
27. Nasharuddin, Nurul Amelina, et al. “Multi-Feature Vegetable Recognition Using Machine Learning Approach on Leaf Images.” International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no. 4, 2019, pp. 1789–94 https://doi.org/10.30534/ijatcse/2019/110842019