PAPER • **OPEN ACCESS**

Utilization of Filter Feature Selection with Support Vector Machine for Tumours Classification

To cite this article: T A H Tengku Mazlin *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **551** 012062

IOP Conf. Series: Materials Science and Engineering **551** (2019) 012062 doi:10.1088/1757-899X/551/1/012062

Utilization of Filter Feature Selection with Support Vector Machine for Tumours Classification

T A H Tengku Mazlin[1], R Sallehuddin[1] and M Y Zuriahati[1]

[1]Faculty of Engineering, School of Computing, Universiti Teknologi Malaysia, 81310, Skudai, Johor Bahru, Malaysia

Email : tgmazlynn@gmail.com, roselina@utm.my, zuriahati@utm.my

Abstract. Due to rapid technology advancement, machine learning has been widely used for solving cancer classification problem. Classification performance is highly depending on the quality of input features. With an explosive increase number of features of high dimensional data, the occurrence of ambiguous samples and data redundancy directly leads to poor classification accuracy. Therefore, this paper presents a utilization of filter feature selection using four filter methods such as Information Gain, Gain Ratio, Chi-Squared and Relief-F by performing attribute rankings to remove the irrelevant and redundant features and evaluate the significance and correlation of input data. Then, the classification will be performed using Support Vector Machine (SVM) to measure the accuracy performance based on the number of selected features. The performance measurement will be validated on standard Breast Cancer datasets consisting of 286 instances obtained from the UCI repository. Evaluation metrics such as accuracy, sensitivity, specificity and Area under Receiver Operating Characteristic Curve (AUC) will be used to assess the performance of the SVM classifier using four different filter methods. Experimental result shows that Gain ratio improves the accuracy of SVM classification compared to Information Gain, Chi-Squared and Relief-F in classifying breast cancer data with only small number of features selected.

1. Introduction

In recent years, the intelligence of machine learning approach such as classification has proven as one of the robust and reliable techniques in increasing the performance of cancer diagnosis. Despite the advantages provided by machine learning, several classification problems in cancer diagnosis have been reported recently. Majority of the classification issues are related to the amount of input variables in the training data or also known as the "curse of dimensionality" [1, 2]. In addition, the classification performance is highly depending on the number of features selected from the training data to train the classifiers [5]. An explosive increase of features in training data has caused a computation complexity where the classification become slower with the presence of noisy, redundant and irrelevant features that consequently degrade the performance of learning tasks [3]. These may cause difficulty to perform data interpretation due to the inconsistency and confusing data patterns. Regarding to the dimensionality issues, some irrelevant and redundant input features which occurred in high dimensional data should be discard in order to maintain or improve the accuracy performance. Hence, the classification model should able to produce an accurate and computationally fasters prediction with optimal number of input variables.

In this work, we present a utilization of four filter feature selection algorithms using Information Gain, Gain Ratio, Chi-square and Relief-F to determine the sets of relevant features. This research

aims to improve the classification accuracy while predicting the optimal number of significant features reliable for classification tasks using SVM classifier. Breast Cancer dataset obtained from UCI Library consisting 286 features were used for evaluating the performance of the proposed utilized method.

The organization of this paper are sorted as follows; section 2 presents the related work. The utilization of four filter feature selection algorithms is explained in section 3 while the implementation of classification algorithm is presented in section 4. All experimental results are discussed in section 5 and section 6 concludes the paper.

2. Related Work

An explosive number of features in high dimensional data have produce challenges towards the classification approach because the learning algorithms are unable to manage the overfitting data which resulting a poor classification accuracy due to the high computational time. As mentioned in reported study, many researchers have developed a classification model with feature selection approach to handle the explosive number of features in the training data for more accurate prediction.

The classification of breast cancer using SVM is proposed in [4] where SVM classifier is implemented with radial basis function (RBF) kernel function to improve the classification performance. In this context, the performance of classifier is highly depending on the quality of the training data which used to train the classifier. For that reasons, it is ultimately important to recognize and identify the doubtful features in the training data at the early phase of the classification. Besides classification techniques, feature selection is identified as one of the techniques used to address the issues. Since choosing the best features could be a challenging task, there are several solutions and methods have been proposed to remove the ambiguous and redundant features of training data.

Feature selection can be separated into three categories namely filter, wrapper and embedded techniques. According to studies on determining the significant features in [5] and [6], filter techniques outperformed wrapper and embedded techniques in terms of input space dimensionality reduction. The filter techniques have successfully improved the accuracy of the classifier and computationally faster when handling the high dimensional input variables. Unlike wrapper techniques, the classification tends to perform inefficiently if the input variables increase rapidly in the input space. As a result, the wrapper technique is more expensive when large input variables need to be used as the computational complexity increased with the dependency of classifiers. Even though the filter feature selection is more efficient when dealing with large dataset, the single filter method does not consider the correlation between each feature, which unable to produce an optimum selection of significant features [5,6]. Despite this, the filter techniques still proven the advantages of computationally faster and ability in handling the high dimensional datasets efficiently and highly recommended due to its simplicity and success performances [3,5,6].

Therefore, the classification in this research will be implemented using SVM classifier where its performance will be improved by utilizing four filter feature selection algorithms and each utilization will be compared with performance parameters such as accuracy, specificity, sensitivity and AUC.

3. Feature Selection using Filter Techniques

Basically, filter techniques are performed without including any learning algorithm or independent from classifiers. This caused the filter techniques to be computationally efficient as the time to compute the algorithm is reduced and ability in handling different dimension of datasets. The selection of features using filter approach is done by assigning a score or rank to each input features. Then, the ranked features are selected into the classifier or eliminated from the dataset according to the scored value. Due to the simplicity and attempts to consider the interaction between features, four filter feature selection algorithms such as Information Gain (IG), Gain Ratio (GR), Chi-Squared (CS) and Relief-F (RF) were implemented for selecting best features. Justification of choosing these four filter algorithms and the equations that implemented can be referred in [7], [8], [9] and [10] accordingly.

3.1 Utilization of Four Filter Feature Selection Algorithms with SVM

Overall, the utilization of four filter feature selection using Information Gain, Gain Ratio, Chi-Squared and Relief-F with SVM is illustrated in figure 1 and the algorithm to perform feature (attribute) ranking is presented as follows:

Step 1: Let X_i be the original feature set of UCI Breast Cancer dataset, where $X_i = \{X_1, X_2, \ldots, X_{286}\}$ and C_i represents the class of benign or malignant, where $C_i = \{C_1, C_2\}$.
Step 2: Rank and sort features X_i for each filter algorithm based on rank value in determining the output class C_i.
Step 3: Select the significant features based on threshold value and output as X'_i. The treshold value used for Information Gain, Gain Ratio, Chi-Squared and Relief-F are entropy value, gain ratio value, significance level and feature score which are set to 0.05 [7,8,9,10].
Step 4: Perform SVM classification on each X'_i and calculate the performance.

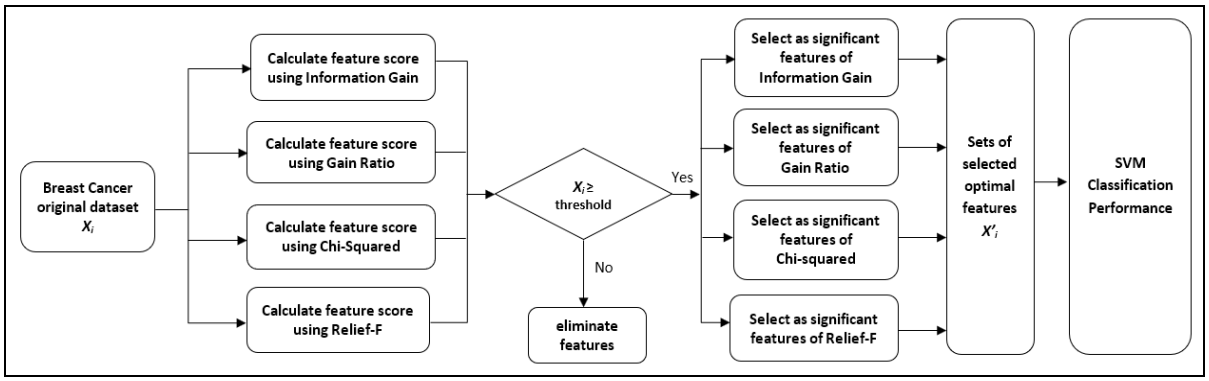

Figure 1. Utilization of four filter feature selection algorithms with SVM.

For the classification performance, SVM using Radial Basis Function (RBF) kernel function will be employed on the selected ranked features of each filter algorithms. Detail implementation of SVM using RBF kernel can be obtained in [11,12]. The SVM performance is measured using four statistical performance namely accuracy, specificity, sensitivity and AUC. Formulas for these statistical performances can be obtained in [2,11,12].

3.2 Experimental Data

UCI Breast Cancer dataset as described in table 1 is used to carry out the experiment. For performance evaluation, the classification dataset needs to be partitioned into training and test sets. Here, 10-fold cross validation is implemented.

Table 1. UCI Breast Cancer original dataset features.

Features	Age	Menopause	Tumour-size	Inv-nodes	Nodes-caps	Deg-malig	Breast	Breast-quad	Irradiat
Attributes	1	2	3	4	5	6	7	8	9
Notation	A1	A2	A3	A4	A5	A6	A7	A8	A9

4. Result and Discussion

This section presents the results of the utilization of four filter feature selection algorithms towards SVM classification. The results of selected features based on average ranking score are described in table 2 and figure 3 summarizes the overall SVM classification performance.

Based on analysis in table 2, four filter algorithms are defined as IG, GR, CS and RF. The information from the top side of the table indicates the features with highest significance and vice versa. For each filter method, the features that obtained values less than threshold value are discarded

from the list as they are considered as not relevant or redundant. This also means that the information provided by these attributes may consists of redundant information which may slows down the SVM learning algorithm in the classifier. For example, Information Gain analysis, the attributes are ranked in descending order such as A6, A4, A3, A5, A9, A1, A8, A7 and A2. Based on the average merit score, attribute A6, A4, A3 and A5 are ranked with the highest average merit score higher than 0.05 among other attributes.

Table 2. Ranking score using Information Gain, Gain Ratio, Chi-Squared and Relief F[a].

	Attribute Ranking				**Average Merit Score** (score with standard deviation (+-))							
	IG	**GR**	**CS**	**RF**	**IG**		**GR**		**CS**		**RF**	
Descending Order	**A6***	**A5***	**A6***	**A6***	0.078	+- 0.011	0.071	+- 0.008	28.875	+- 4.215	0.093	+- 0.014
	A4*	**A4***	**A4***	**A8***	0.071	+- 0.01	0.054	+- 0.007	26.594	+- 3.857	0.062	+- 0.013
	A3*	**A6***	**A5***	**A2***	0.061	+- 0.008	0.051	+- 0.007	19.731	+- 2.515	0.057	+- 0.009
	A5*	A9	**A3***	**A1***	0.051	+- 0.007	0.033	+- 0.007	17.039	+- 1.913	0.051	+- 0.011
	A9	A3	**A9***	A7	0.026	+- 0.006	0.02	+- 0.003	9.792	+- 2.248	0.048	+- 0.005
	A1	A1	**A1***	A3	0.012	+- 0.003	0.006	+- 0.002	3.956	+- 1.339	0.005	+- 0.007
	A8	A8	A8	A9	0.01	+- 0.003	0.005	+- 0.002	3.462	+- 1.183	0.033	+- 0.008
	A7	A7	A7	A5	0.003	+- 0.001	0.002	+- 0.002	0.887	+- 0.6	0.026	+- 0.005
	A2	A2	A2	A4	0.003	+- 0.002	0.003	+- 0.001	0.94	+- 0.437	0.018	+- 0.01

[a] Hint: **X*** indicates the selected features (attributes).

From the analysis, we observed that certain features are similar on different filter techniques even though different ranking methods were applied in each filter. From the ranking score analysis, 4 features are selected by Information Gain, 3 features are selected by Gain Ratio, 6 features are selected by Chi-Squared and 4 features are selected by Relief-F. As an achievement from Algorithm 3.1, these selected significant features are the optimum features to be included into SVM for classification tasks.

4.1. Performance of SVM Classification based on Selected Features

The performance measures such as accuracy, sensitivity, specificity, and AUC are evaluated and averaged to obtain the overall SVM performance. Figure 2 presents the overall results obtained.

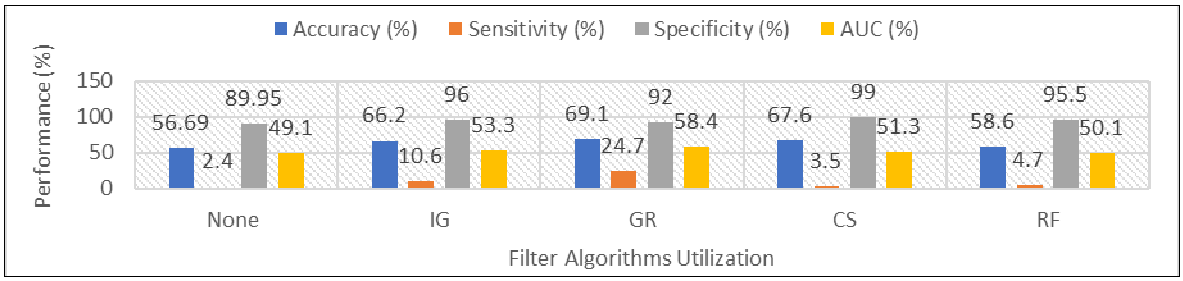

Figure 2. Overall SVM classification performance with four filter algorithms utilization.

Based on the results in figure 2, the classification performance achieved with feature selection utilization is better compared to the independent classifier. For example, the performance of SVM with Gain Ratio utilization perform better than others in terms of accuracy and sensitivity. However, in terms of specificity, SVM classification with Chi-Squared utilization achieved higher specificity (99.00%) compared to the utilization with Information Gain (96.00%), Gain Ratio (92.00%) and Relief-F (95.50%). AUC measures the effectiveness of classifier in terms of correctly classified the true positive (benign) and true negative (malignant) classes. It is observed that the AUC value of SVM with Gain Ratio utilization is better with highest percentage (58.40%) compared to the utilization with Information Gain (53.30%), Chi-Squared (51.30%) and Relief-F (50.10%). Based on the AUC performance obtained by SVM with four filter feature selection algorithms, SVM classification with

Gain Ratio utilization could identify the class of the data accurately than other three filter algorithms. By selecting 3 significant features (A5, A4, A6), the classification performance with Gain Ratio shows an improvement in terms of accuracy. This indicates that the other 6 features contain some redundancy, which have less contribution towards determining a certain class (benign) and (malignant). For overall effectiveness, it can be proved that filter feature selection using Gain Ratio can significantly improves the accuracy of SVM classification compared to Information Gain, Chi-Squared and Relief-F in classifying breast cancer data with only small number of features selected.

5. Conclusions

Handling massive increment of ambiguous and redundant samples in the training data has resulting a notable challenge faced by the current cancer diagnosis system in machine learning. Feature selection method is one of the solutions used to pre-process the datasets before proceeding it into the cancer classification process. In this paper, the performance of utilization of feature selection using four filter algorithms such as Information Gain, Gain Ratio, Chi-Squared and Relief-F has been examined on UCI Breast Cancer dataset to determine the significant features by ranking of attributes. Sets of threshold value are used to select the significant features from the ranked sample.

Experimental results show that SVM classification with gain ratio feature selection utilization contributes better performance than Information Gain, Chi-Squared and Relief-F utilization for breast cancer classification in terms of accuracy, sensitivity and AUC with percentage of 69.10%, 24.70% and 58.40%. This research indicates that SVM classification with gain ratio feature selection utilization could effectively improves the accuracy as well as helping medical experts in diagnosing breast cancer. For the future work, this research will be continued to optimize the number of significant features for optimal solution with different optimization approach while increasing the accuracy performance.

Acknowledgments

This research is supported by a Research University Grant, GUP TIER 1:16H57 and GUP TIER 2:14J13. The authors would like to thank the anonymous reviewer for providing constructive and generous feedback. The authors also would like to thank Research Management Centre (RMC) of Universiti Teknologi Malaysia, for the research activities and Applied Industrial Analytics (Ali@s) for the support and motivation in making this research a success.

References

[1] Wu J *et al* 2017 *J. of Biostatistics and Biometrics* **1(2)** 1-7
[2] Sharifah H *et al* 2013 *J. of Technology Science & Engineering* **65(1)** 73-81
[3] Miao *et al* 2016 *J. of Proc. Computer Science* **91** 919–926
[4] Liu B *et al* 2010 *J. of Pattern Recognition* **43(1)** 280–298
[5] Wang D *et al* 2018 *J. of Computing and Applied Mathematics* **329** 307- 321
[6] Sharifah H *et al* 2014 *Classification of Liver Cancer Using Artificial Neural Network and Support Vector Machine in Proc.of Int. Conf. on Adv. in Communication, Network and Computing CNC* February 21-22 Chennai, India pp 1-6
[7] Tresna M *et al* 2017 *Int. J. of Engineering Technology EMITTER* **5(1)** 36-71
[8] Dai J *et al* 2013 *J. of Appl. Soft Computing* **13** 211–221
[9] Seema A *et al* 2018 *Int. J. of Engineering and Technology* **4(2)** 222-226
[10] Ryan J U et al 2018 *J. of Biomedical Informatics* **85** 168–188
[11] Sallehuddin R *et al* 2016 *J. of Technology Science & Engineering* **78** 107–119
[12] Huang S *et al* 2018 *J. of Cancer Genomics & Proteomics* **15(1)** 41–51