Int. J. Advance Soft Compu. Appl, Vol. 10, No. 1, March 2018
ISSN 2074-8523

Filter-Wrapper Combination and Embedded Feature Selection for Gene Expression Data

Shilan S. Hameed[1,2], Olutomilayo Olayemi Petinrin[1], Abdirahman Osman Hashi[1], and Faisal Saeed[1,3*]

[1]Information Systems Department, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia.
[2]Department of Software and Informatics Engineering, College of Engineering, Salahaddin University, Erbil, Kurdistan Region, Iraq.
[3]College of Computer Science and Engineering, University of Taibah, Medina, Saudi Arabia
Email: shilansamin@gmail.com, olutomilayo.petinrin@gmail.com, wadani12727@gmail.com, fsaeed@taibahu.edu.sa

Abstract

Biomedical and bioinformatics datasets are generally large in terms of their number of features - and include redundant and irrelevant features, which affect the effectiveness and efficiency of classification of these datasets. Several different features selection methods have been utilised in various fields, including bioinformatics, to reduce the number of features. This study utilised Filter-Wrapper combination and embedded (LASSO) feature selection methods on both high and low dimensional datasets before classification was performed. The results illustrate that the combination of filter and wrapper feature selection to create a hybrid form of feature selection provides better performance than using filter only. In addition, LASSO performed better on high dimensional data.

Keywords: *Bioinformatics; Gene Expression Data; Feature Selection; Filter Methods; Wrapper Methods.*

1 Introduction

A feature is a stand-alone characteristic of an instance being observed that can be readily measured. A set of features can be used in machine learning algorithm for classification [1]. Feature selection has been a growing and developing research field since the 1970's and has proven productive for removing redundant and irrelevant features, thereby increasing learning task efficiency and the predictive performance of learning methods, as well as improving abilities for comprehending the results from learning methods [2]. 'Feature selection' is a term commonly used in machine learning and statistics. It involves the selection of a subset of relevant features, before the construction of a model. Feature selection has been used on various ranges of data, including both low and high dimensional data. With regard to high dimensional data it is basically used to remove redundant and unwanted features. There is a clear and recognised need for feature selection techniques with regard to several applications of bioinformatics [3]. Biomedical and bioinformatics related data usually have a large number of input features and are characterized by high dimensionality that can significantly increase the computational burden [4].While there are cases which have low numbers of features, these are usually rare. The redundant features do not contribute to modelling a better predictor, since the information they provide is basically presented by other feature(s) [5]. It is imperative to know that redundant features negatively affect the performance of a model, and in order to achieve better performance, it is desirable to perform feature selection. Furthermore, features which are irrelevant do not only negatively affect the accuracy of classifiers, but also create added difficulties when searching for useful knowledge [3]. The exclusion of irrelevant features facilitates the visualization of data and hence makes the computational models more easily understood. Therefore, feature selection, a concept whose purpose is the finding of a subset of discriminative features, becomes essential, and is widely recognised as one of the centrally important areas in biomedical and bioinformatics data mining [6]. It is worth considering the cost minimization of database storage and management that occurs through feature selection, as it reduces the initially required measurement and storage [2]. Genetic data are found largely in microarray databases. When such data are properly analysed, the comprehension and understanding of medicine and biology can be improved. The genetic mechanism of proteins and cancers can be inspected by carrying out several microarray experiments and, over time, systematic approaches have been utilized for both the classification of different cancer types and to differentiate between noncancerous and cancerous tissues, as well as identifying protein structures [7]. Over the last decade, machine learning methods have been utilized in the analysis of microarray data. Various approaches have been implemented in order to: (i) classify several cancer types; (ii) distinguish between noncancerous and cancerous samples and; (iii) identify the aggressive progression of some subtypes of cancer. These analyses are all targeted towards the generation of interpretations from complex datasets which

are biologically meaningful and thus to suggest an experimental follow-up. Microarray analysis is utilized for finding discriminating biomarkers which assists in the identification of tissue types and which has wide application in cancer studies. However, the classification of data samples in microarray data is not an easy task due to the vast number of genes involved, which goes up to the tens of thousands, and the inversely low number of samples which amount to hundreds [8] . Feature selection can be used to tackle this problem, enabling the most informative genes to be discovered. The basic feature selection types are; filter, wrapper, embedded, and hybrid [2, 3, 7].

The filter selection method chooses variables regardless of the model used. This method is based only on the general features, such as the association with the variable, to predict. The filter methods work by suppressing variables that are least interesting. The non-suppressed variables will be a part of a regression or a classification model which is used for the classification or prediction of data. Filter methods are robust in terms of overfitting and showing effectiveness in computation time [9]. As a general rule, these methods estimate a relevance score, while a threshold scheme is used to select the best-scoring features/genes. Filter techniques are not necessarily used to build predictors [10]. Taking the distributed data into consideration, filters can be categorized as among parametric and non-parametric methods. Parametric filters assume equal distribution of samples in different classes, such ANOVA , chi-squared and Bayesian [3]. However, this assumption cannot be guaranteed in most datasets. Therefore, the utilization of non-parametric methods might yield a better result when there is uncertainty regarding the dataset distribution. Examples of non-parametric filters are ReliefF, Information gain and Gain ratio. In the wrapper based feature selection, the evaluation is performed on subsets of the variables, through which, unlike with the use of filter methods, the possible communications between the variables can be observed. This is achieved by using the classifier accuracy [3]. Wrappers choose the best subset of features that gives highest accuracy to the model. The result of this selection usually consists of fewer number of features with robust discriminative power [11]. In addition, wrappers are classifier dependent, and hence the same result is not guaranteed when another classifier is applied [10, 12]. Therefore, whenever a wrapper method is used, it is recommended that different classifiers are applied for the feature selection. The third type of feature selection is embedded methods. This is similar to wrapper approaches, in that they are dependent on a given learning algorithm. Nevertheless, these methods can interact with the classifier, while being less computationally intensive than wrapper methods [3, 10]. Hence, embedded methods are expected to combine the efficiency of filters with the accuracy of wrappers. They are implemented in such a manner that their built-in feature selection is performed by the reduction of features. LASSO and RIDGE regression are two major examples of the embedded method [2, 3]. In this paper, we investigate the performance of these three features selection methods for biological datasets.

2 Related Work

Various feature selection methods have been used by researchers for protein and gene related dataset selection and classification. These methods have been proven to be effective, though none of them have been shown to be the best method available. In previous studies, different kinds of feature selections were applied - including filters and wrappers, as well as the combination of the two. Researchers [13] have applied the combination of four filter methods, namely; Information Gain, x2, Odds-Ratio, and Correlation Coefficient with Genetic Programming (GP), in order to gain the advantages provided by the different metrics. Doing so increased the efficiency, as well as providing higher accuracy, through an aggressive reduction in features. However, applying GP directly to high dimensional data is computationally expensive. In another study [14], a combination of information theory and LASSO was utilised for feature selection. Their analysis showed that the proposed method was effective. Another study [15] used entropy gain and neighbourhood roughest to overcome the deexcitation problems of roughest when used for gene selection, as well as to improve the accuracy of classification. In other research [16], a family of embedded methods for backward feature selection using SVM was utilized. Their method showed better accuracy with regard to four out of six databases. A further study [17] utilized a hybrid form of filter and wrapper, consisting of information gain and standard genetic algorithm. In other research [18], a bid to tackle the issue of time consumption, cost, and efficiency while investigating protein post-translational modification, utilised Dagging method as a classification technique in the prediction of N-formylated methionine. Minimum Redundancy, Maximum Relevance (mRMR) and Incremental Feature Selection (IFS) were used as the feature selection methods. mRMR is used for ranking features, from the most important to the least important, and IFS is used in the selection of the optimal features from amongst the ranked features. The model which adopted the optimized features performed best, with an accuracy of 90.74% and MCC value of 0.7478. The same method has also been used for the classification of protein domain [19], and it was proposed that this would be a more useful complement to methods like DoMpro, Globplot, and Domcut. Jia and Du [20] also used minimal redundancy, maximal relevance (mRMR) as a method of feature selection in the prediction of Golgi-resident protein by discretizing the features to (-1, 0, +1). Ahmad et al. [21] on the other hand, utilised the Fisher selection method to reduce noise and redundant features, before the classification of sub-Golgi protein. Yuan et al. [22] maximized correlation information using a method known as Maximum Correlation Information (MCI). This method evaluated the importance of each feature by maximizing the correlation information between the class coding space and the feature space. The main contribution of the current research work is to explore the application of a combination of feature selection methods based on filter, wrapper, and embedded approaches and to give a comparative analysis

using six well-known gene expression datasets, which include the low and high dimensional structure.

3 Methodology

3.1 Datasets

Six datasets (three high dimensional datasets and three low dimensional datasets) were analysed in this research. The high dimensional datasets are a leukaemia cancer dataset, a colon cancer dataset, and prostate cancer microarray datasets. The dataset of leukaemia was originally presented in Golub et al. [23]. The dataset was generated from a gene expression study in two types of acute leukaemia: acute myeloid leukaemia (AML) and, acute lymphoblastic leukaemia (ALL). The levels of gene expression were measured using Affymetrix high-density oligonucleotide arrays which consist of 6817 genes - although this was reduced to 3051 genes and further analysed by Golub et al. [23]. The data consist of 25 cases of AML and, 47 cases of ALL (38 B-cell ALL and 9 T-cell ALL). The dataset was further pre-processed by Dudoit et al. [24]. The colon cancer microarray data set was originally analysed by Alon et al. [25]. The original authors of the data set performed treatment on the raw data from the Affymetrix oligonucleotide arrays. The intensity of the full-length gene of a particular array is divided by the mean intensity of all the full-length gene on the same array and multiplied by an average nominal intensity of 50, in order to offset all the possible variations which can exist between the arrays. The dataset is a binary class, consisting of normal and tumour tissue samples. The total number of samples are 62 and total gene numbers after pre-processing by previous authors is 2000. This microarray data set was originally analysed by Singh et al. [26]. The prostate cancer dataset consists of 102 patterns of gene expression, where 50 of the samples are normal prostate specimens and the other 52 are tumours. The dataset, which is a gene expression data, is based on oligonucleotide microarray, and consists of approximately 12600 genes. After pre-processing the remaining number of genes in the data set is 6033. The low dimensional dataset are Yeast Protein localization sites dataset, E-coli protein localization sites dataset, and Mice protein expression dataset, which were retrieved from the UCI database repository. The Yeast dataset contains 1484 instances and 9 attributes. The E-coli dataset contains 336 instances and 8 attributes. The Mice protein dataset contains 1071 instances and 9 attributes. These datasets are all bioinformatics related datasets of protein and gene expression. These datasets which contain missing values are pre-processed and the data have been normalized as well.

Table 1: The six high and low dimensional dataset characteristics

Datasets	# features	# samples	type	# classes
Colon	2000	62	High dimensional	2 (22- 40)
Prostate	6033	102	High dimensional	2 (50-52)
Leukaemia	3051	72	High dimensional	2 (47-25)
Mice Protein	82	1080	Low dimensional	8 classes
Yeast	8	1484	Low dimensional	10 classes
E. coli	7	336	Low dimensional	8 classes

3.2 Applied feature selection methods

Three features selection methods were implemented in this study. These methods are ReliefF filter, Wrapper, and LASSO (Least Absolute Shrinkage and Selection Operator). The first two methods were applied using Weka machine learning tool packages [27] while Rapid miner was used for the third.. Filter feature selections are generally applicable for handling high dimensional datasets, in which the feature numbers are considerably larger than the sample numbers. However, the results of such methods are not sufficiently accurate and robust to depend on, because they do not receive assessment feedback from the applied model [12]. Consequently, their use as a pre-processing route, before the final selection is being made, could be the best possible approach [28]. In this study, ReliefF was employed, due to its unique performance in bioinformatics applications [9]. Further, it was observed in the current study that ReliefF showed better performance compared to that of the other filters. The ReliefF filter [29] acts by ranking the features according to their highest correlation with a specific class, while their distance to the other class is also taken into account [30]. This method has a number of advantages, such as its independency on heuristics, its short running time, and the fact that it is powerful against noise. If a dataset belonging to a binary class with n samples and f features for each sample is assumed, the algorithm takes t number of iterations. Initially, a zero-filled f-length weight vector (W) is generated. For each iteration, the feature vector (x) of a randomly selected instance was compared with the feature vector of a closest instance to it (using Euclidian distance) from both of the classes. The same class instance having closest distance is described as 'near-hit', while the opposite class nearest instance is described as 'near-miss' [31]. The weight vector was updated as follow;

$$\{W_i = W_i - (x_i - nearHit_i)^2 + (x_i - nearMiss_i)^2 \quad [1]$$

The result of the weight vector for each feature after the total number of iterations was divided by the *t* number of iterations. This value is ranked according to the *p*-value or a threshold value.

WrapperSubsetEval is a default wrapper method in Weka tool which was implemented using greedy stepwise search in this study. It is a well-known form of wrapper feature selections search method which involves adding or removing the features based on their discriminative powers [32]. In the current work, the forward feature selection scheme was used for feature selection in the greedy stepwise mode. LASSO is a feature selection which is embedded in a classifier. It is a powerful method that performs two main tasks: regularization and feature selection. The LASSO method puts a constraint on the sum of the absolute values of the model parameters; the sum has to be less than a fixed value (upper bound). In order to do so the method applies a shrinking (regularization) process, where it penalizes the coefficients of the regression variables, shrinking some of them to zero.

3.3 Experimental Design

During the features selection process, the variables that still have a non-zero coefficient after the shrinking process are selected to be part of the model, with the goal of minimizing the prediction error. Prediction analysis is carried out on all the datasets using classifiers which are commonly used in machine learning predictive analysis. These classifiers are: Bayes Net, Support Vector Machine, Naïve Bayes, and k-Nearest Neighbour. These classifiers were evaluated based on the accuracy, Sensitivity, Specificity and G-mean both before and after feature selection was conducted on the datasets. Wrapper and LASSO were used for the low dimensional datasets in two different applications, while for the high dimensional data, RreliefF filter is used as first step then LASSO and wrapper are applied on its result, again in two separate applications. The RreliefF filter was used in ranking the attributes according to their importance in 10-fold cross validation form. Subsequently, the first 100 high ranked attributes were used in classification. LASSO and wrapper were then separately applied on the filtered dataset of 100 attributes. The experimental design of this research is shown in Fig. 1. This study implemented feature selection for each classifier separately, in order to avoid biased results. Moreover, when there are an insufficient number of observations, there would be a risk of overfitting, whereas using wrappers on a large number of variables causes an increase in the computational time necessary [33]. For these reasons, the high dimensional datasets were first filtered and then the wrapper method was applied.

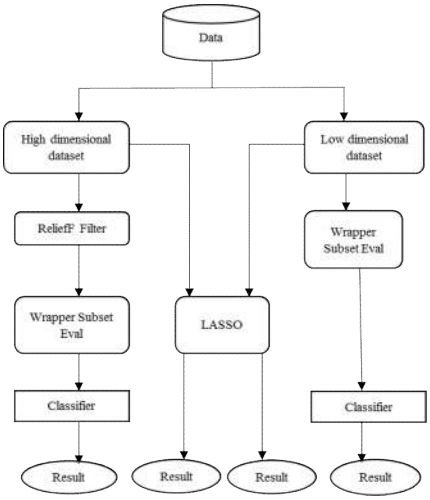

Fig. 1: Experimental Design

4 Results, Analysis and Discussions

The accuracy of the classifiers with regard to the original high and low dimensional datasets were evaluated first. A 10-fold cross validation was applied with each classifier for the training and testing. The results in Table 2 illustrate that although Support Vector Machine presented the highest classification accuracy amongst all the classifiers, it does not provide better performance on low dimensional datasets.

Table 2: Accuracy of Classifiers with Original Dataset

Dataset	Classifier			
	Bayes Net (%)	Naïve Bayes (%)	k-Nearest Neighbour (%)	Support Vector Machine (%)
Prostate	82.35	62.75	85.29	**91.18**
Colon	**88.71**	83.87	77.58	**88.71**
Leukaemia	97.22	97.22	97.22	**98.61**
Mice Protein	94.35	87.50	99.26	**100.00**
Yeast	56.74	**57.61**	52.29	57.01
E-coli	82.09	**85.37**	80.30	84.48

The features of the three high dimensional datasets were ranked using ReliefF filter feature selection method using 10-fold cross validation. Thus, the features are ordered based on the ranking results. To avoid overfitting in the next steps of feature selections (wrapper), possible due to having low number of samples, the first 100 important attributes were selected. Subsequently. these features were used by the four classifiers. The results of the classifiers performance are tabulated in Table 3-a and 3-b. Results in bold indicate the best accuracy for each

dataset using the same classifier. Results highlighted in grey, show the best selected approach for each dataset. The dashed cells indicate that the method is not appropriate for application. The results show that generally the accuracy of the classifiers on the filtered dataset illustrate better results, when compared with the one on original datasets. However, for the leukaemia dataset this was true only for the BN classifier. It was noticed that the features of the low dimensional datasets were not selected by the filter since it has a low number of initial features, and that using the filter gives discrimination ability to all features.

In the next step, wrapper feature selection was applied to the six datasets. For the high dimensional datasets, it was applied to the reduced dataset with 100 attributes that were selected by the ReliefF filter method, which is considered a hybrid method. Table 3-a illustrates the better performance of the hybrid feature selection method (combined filter and wrapper) on the high dimensional datasets. The accuracy of this hybrid method was better (in 10 out of 12 cases) when compared with the accuracy of the same classifiers applied on original and filtered datasets.

Table 3-a: classifiers accuracy applied on high and low dimensional datasets before and after application of filter, wrapper, and LASSO.

Dataset	Status	Bayes Net (%)	Naïve Bayes (%)	k-Nearest Neighbour (%)	Support Vector Machine (%)	Lasso (%)
Prostate	original	82.35	62.75	85.29	91.18	**96.09**
	filtered	**94.12**	93.14	91.18	93.14	96.09
	wrapper	93.14	**97.06**	**95.10**	**94.12**	-
Colon	original	88.71	83.87	77.58	88.71	**93.81**
	filtered	87.10	88.71	82.26	88.71	87.14
	wrapper	**91.94**	**90.32**	**87.10**	**91.94**	-
Leukaemia	original	97.22	97.22	**97.22**	**98.61**	96.07
	filtered	98.61	95.83	95.83	97.22	94.64
	wrapper	**100.00**	**97.22**	84.72	**98.61**	-
Mice Protein	original	94.35	87.50	99.26	**100.00**	99.17
	filtered	-	-	-	-	-
	wrapper	**100.00**	**100.00**	**100.00**	**100.00**	-
Yeast	original	56.74	57.62	**52.29**	**57.01**	56.94
	filtered	-	-	-	-	-
	wrapper	56.74	**57.88**	52.16	56.67	-
E-coli	original	82.09	**85.37**	80.30	**84.48**	81.22
	filtered	-	-	-	-	-
	wrapper	82.09	**85.37**	**81.20**	83.88	-

As mentioned earlier, wrapper feature selection was applied on the low dimensional dataset without any prior feature selection. By comparing this result with the accuracy of classifiers on the original data, as shown in Table 3-a, it can be noticed that the accuracy of classifiers on the datasets after using wrapper feature selection was better for the Mice Protein dataset, which showed better

result for all classifiers. However, for the other two low dimensional datasets, improvements have been obtained with regard to some classifiers only.

However, this study used some datasets that contain an imbalanced number of samples for binary class datasets (the Colon and Leukaemia datasets). To draw a better conclusion of the model performance, it may not be sufficient to rely on accuracy alone. For this reason, we have used sensitivity, specificity as well as G-mean (the square root of sensitivity multiplied by specificity) of the achieved performance as presented in Table 3-b for high dimensional datasets and Table 3-c for low dimensional datasets. The results in bold represent the best sensitivity, specificity and G-mean for each dataset using the same classifier. In addition, for each dataset, results highlighted in grey show the best selected approaches for each of the three measures: sensitivity, specificity and G-mean.

Table 3-b illustrates that for imbalanced, high dimensional datasets the models' performance on the original and filtered datasets are generally affected by the skewed samples in terms of specificity and sensitivity. However, this effect was not sufficiently significant. Moreover, the imbalance in the results between sensitivity and specificity is not pronounced after the application of the wrapper feature selection models. This indicates that the wrapper application is better when compared with other models.

Table 3-b: classifiers sensitivity, specificity and G-mean applied on high and low dimensional datasets before and after the application of filter, wrapper, and LASSO.

Dataset	Status	Bayes Net (%)			Naïve Bayes (%)			k-Nearest Neighbour (%)			Support Vector Machine (%)			Lasso (%)		
		Spec.	Sens.	G-mean	Spec.	Sens.	G-mean	Spec.	Sens.	G-mean	Spec.	Sens.	G-mean	Spec.	Sen.	G-mean
Prostate	original	84.0	80.8	82.4	50.0	75.0	61.2	84.0	86.5	85.2	92.0	90.4	91.0	**100.0**	92.3	**96.1**
	filtered	**98.0**	90.4	**94.1**	96.0	90.4	93.2	94.0	88.5	91.2	96.0	90.4	93.2	98.0	**94.0**	96.0
	wrapper	94.0	**92.3**	93.0	**98.0**	**96.2**	**97.1**	**96.0**	**94.2**	**95.1**	**98.0**	**90.4**	**94.1**	-	-	-
Colon	original	86.4	90.0	88.2	77.3	87.5	82.2	68.2	75.0	71.5	86.4	90.0	88.2	**95.5**	**92.5**	**94.0**
	filtered	86.4	87.5	86.9	86.4	90.0	88.2	68.2	90.0	78.3	86.4	90.0	88.2	81.8	90.0	85.8
	wrapper	**95.5**	**90.0**	**92.7**	**90.9**	**90.0**	**90.4**	**81.8**	**90.0**	**85.3**	**86.4**	**95.0**	**90.6**	-	-	-
Leukaemia	original	97.9	96.0	96.9	**100.0**	92.0	95.9	95.7	**100.0**	**97.8**	100.0	96.0	98.0	100.0		93.8
	filtered	97.9	100.0	98.9	95.7	96.0	95.8	**97.9**	92.0	94.9	97.9	96.0	96.9	**100.0**	84.0	91.7
	wrapper	**100.0**	**100.0**	**100.0**	97.9	**96.0**	**96.9**	89.4	76.0	82.4	**100.0**	**96.0**	**98.0**	-	-	-

Taking sensitivity and specificity into account individually, we can see that there are few models that possess high specificity or sensitivity only, while the wrapper model presented high sensitivity and specificity among all the datasets. As mentioned previously, the bold value in the Table 3-b indicates improvement in sensitivity and specificity for each classifier, while those highlighted in grey indicated the best approach among all classifiers for each dataset.

Furthermore, in terms of G-mean, if we compare the results with the accuracy of the same classifier, we can see that the results are quite similar to each other, with only a slight difference in some classifiers. This implies satisfaction with regard to most of the models' accuracy. The highest value of G-mean is once again obtained in the wrapper applications.

The sensitivity and specificity of the low dimensional datasets are not included in Table 3-c due to the high number of classes of such datasets. Nevertheless, they are balanced datasets. For this reason, we depend on the value of G-mean for comparison. It can be seen that the G-mean value of the low dimensional datasets are also similar with regard to their accuracy. This indicates that the classifiers performance is not biased.

Table 3-c: classifiers G-mean applied on low dimensional datasets before and after the application of wrapper, and LASSO.

Datasets	Status	Bayes Net (%)	Naïve Bayes (%)	k-Nearest Neighbour (%)	Support Vector Machine (%)	Lasso (%)
Mice Protein	original	94.4	87.5	99.3	100.0	99.5
	wrapper	**100.0**	**100.0**	**100.0**	**100.0**	-
Yeast	original	56.7	57.6	52.2	57.0	56.0
	wrapper	56.7	**57.9**	52.2	56.7	-
E-coli	original	82.1	85.4	80.3	**84.5**	81.9
	wrapper	82.1	**85.4**	**81.2**	83.9	-

It is worth mentioning that in the wrapper feature selection of high dimensional datasets, most of the classifiers selected different number of features, having almost similar features(genes) among them. The results of the reduced dataset after the application of the wrapper are tabulated in Table 4. One can see that all of the classifiers selected a very low number of genes with highest accuracy. The best approach for wrapper application is to achieve multi-objective feature selection, in which the highest accuracy with few subsets of features can be obtained, especially in the cancer informatics.

Table 4: The effect of filter and wrapper on the number of selected features by the applied classifiers in all the datasets.

		# Features after filter	# Features in the best subset after wrapper			
Dataset	# Original features	RliefF	Bayes Net	Naïve Bayes	k-Nearest Neighbour	Support Vector Machine
Colon	2000	100	3	4	8	2
Prostate	6033	100	3	3	5	2
Leukaemia	3051	100	3	5	2	9
Mice Protein	82	-	4	5	6	5
Yeast	8	-	7	7	7	7
E-coli	7	-	6	6	5	5

The LASSO feature selection method was also applied to both high and low dimensional datasets and the results are shown in Table 3-a. It was also applied to the original and filtered high dimensional dataset, and to the original low dimensional dataset. It can be seen that the performance of LASSO on the high dimensional data is better than its performance on the low dimensional data. This is because LASSO is a dynamic feature selection tool which uses regression analysis and as such, might be inappropriate for low dimensional datasets. A further insight provided by the results given in Table 3-a, is the indication that LASSO performed better on high dimensional datasets without prior feature selection, than on those with prior feature selection (filter methods). This is because LASSO combines the ability of filter and wrapper feature selections.

Generally, the classification methods showed a competitive performance for all applications. It is a rule of thumb that none of the machine learning methods do well for all datasets and not all of them illustrate similar performances on the same dataset. Nevertheless, from examining the results, it can be suggested that K-nearest neighbour did not perform well when compared with the others, while the rest showed similar performances.

Additionally, in terms of the high dimensional datasets (Leukaemia, Colon, Prostate), when compared with previous studies which were undertaken using only classification [34, 35], the results of this study showed better accuracy after using filter and wrapper, as well as the application of LASSO for the high dimensional data set. It can be concluded that for the selected classifiers, the application of combination of wrapper with filter and LASSO compete well in terms of the classification accuracy.

5 Conclusion

Bioinformatics data are typically datasets with a large number of attributes. Feature selection has been used on various ranges of data, including both low and high dimensional data, although it is primarily utilized with high dimensional data to remove redundant and unwanted features. This paper implemented a range of different feature selection - including ReliefFfilter, Wrapper subsetEval with greedy search, and LASSO as embedded feature selection. Filter was applied individually and incombination with wrapper for high dimensional data. Wrapper and LASSO were applied directly to low dimensional data. LASSO was applied to high dimensional data for original and filtered data. Regarding the high dimensional data, it was illustrated that the combination of filter and wrapper performs better, when compared with applying classification directly on the original and filtered data, in terms of accuracy, balanced sensitivity with specificity and efficiency. The application of LASSO showed competitive performance with the application of combination of filter and wrapper for high dimensional data. However, for low dimensional data, performing only wrapper on original dataset showed better performance when compared with other methods.

ACKNOWLEDGEMENTS

This work is supported by the Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under the Research University Grant Category (VOT Q. J130000.2528.16H74).

References

[1] Chandrashekar, G., Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering 40*(1) 16-28.

[2] Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research 3*(Mar) 1157-1182.

[3] Saeys, Y., Inza, I., Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics 23*(19) 2507-2517.

[4] Wang, L., 2012, Feature selection in bioinformatics, SPIE Defense, Security, and Sensing, International Society for Optics and Photonics, pp. 840113-840113-6.

[5] Song, Q., Ni, J., Wang, G. (2013). A fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE transactions on knowledge and data engineering 25*(1) 1-14.

[6] Conilione, P., Wang, D. (2005). A comparative study on feature selection for E. coli promoter recognition. *Int. J. Inf. Technol 11* 54-66.

[7] Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J.M., Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences 282* 111-135.

[8] Carlos J. Alonso-González, Q.I.M.-S., Arancha Simon-Hurtado, Ricardo Varela-Arrabal. (2012). Microarray gene expression classification with few genes: Criteria to combine

attribute selection and classification methods. *Expert Systems with Applications*.

[9] Hira, Z.M., Gillies, D.F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics 2015*.

[10] Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 9*(4) 1106-1119.

[11] Xiong, M., Fang, X., Zhao, J. (2001). Biomarker identification by feature wrappers. *Genome Research 11*(11) 1878-1887.

[12] Santana, L.E.A.d.S., de Paula Canuto, A.M. (2014). Filter-based optimization techniques for selection of feature subsets in ensemble systems. *Expert Systems with Applications 41*(4) 1622-1631.

[13] Sandin, I., Andrade, G., Viegas, F., Madeira, D., Rocha, L., Salles, T., Gonçalves, M., 2012, Aggressive and effective feature selection using genetic programming, 2012 IEEE Congress on Evolutionary Computation, IEEE, pp. 1-8.

[14] Zhongxin, W., Gang, S., Jing, Z., Jia, Z. (2016). Feature Selection Algorithm Based on Mutual Information and Lasso for Microarray Data. *Open Biotechnology Journal 10* 278-286.

[15] Chen, Y., Zhang, Z., Zheng, J., Ma, Y., Xue, Y. (2017). Gene selection for tumor classification using neighborhood rough sets and entropy measures. *Journal of Biomedical Informatics 67* 59-68.

[16] Salem, H., Attiya, G., El-Fishawy, N. (2017). Classification of human cancer diseases by gene expression profiles. *Applied Soft Computing 50* 124-134.

[17] Maldonado, S., Weber, R., Famili, F. (2014). Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines. *Information Sciences 286* 228-246.

[18] Zhou, Y., Huang, T., Huang, G., Zhang, N., Kong, X., Cai, Y.-D. (2016). Prediction of protein N-formylation and comparison with N-acetylation based on a feature selection method. *Neurocomputing 217* 53-62.

[19] Li, B.-Q., Hu, L.-L., Chen, L., Feng, K.-Y., Cai, Y.-D., Chou, K.-C. (2012). Prediction of protein domain with mRMR feature selection and analysis. *PLoS One 7*(6) e39308.

[20] Jiao, Y.-S., Du, P.-F. (2016). Prediction of Golgi-resident protein types using general form of Chou's pseudo-amino acid compositions: Approaches with minimal redundancy maximal relevance feature selection. *Journal of theoretical biology 402* 38-44.

[21] Ahmad, J., Javed, F., Hayat, M. (2017). Intelligent computational model for classification of sub-Golgi protein using oversampling and fisher feature selection methods. *Artificial Intelligence in Medicine*.

[22] Yuan, M., Yang, Z., Huang, G., Ji, G. (2017). Feature selection by maximizing correlation information for integrated high-dimensional protein data. *Pattern Recognition Letters 92* 17-24.

[23] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science 286*(5439) 531-537.

[24] Dudoit, S., Fridlyand, J., Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association 97*(457) 77-87.

[25] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences 96*(12) 6745-6750.

[26] Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell 1*(2) 203-209.

[27] I.H. Witten, E.F., Mark A. Hall, and Chris J. Pal. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". *Fourth Edition*.

[28] Das, A.K., Das, S., Ghosh, A. (2017). Ensemble feature selection using bi-objective genetic algorithm. *Knowledge-Based Systems 123* 116-127.

[29] Kira, K., Rendell, L.A., 1992, The feature selection problem: Traditional methods and a new algorithm, Aaai, pp. 129-134.

[30] Robnik-Šikonja, M., Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning 53*(1-2) 23-69.

[31] Kwak, N., Choi, C.-H. (2002). Input feature selection by mutual information based on Parzen window. *IEEE transactions on pattern analysis and machine intelligence 24*(12) 1667-1671.

[32] Freitag, D., 2017, Greedy attribute selection, Machine Learning Proceedings 1994: Proceedings of the Eighth International Conference, Morgan Kaufmann, p. 28.

[33] Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and information systems 34*(3) 483-519.

[34] Díaz-Uriarte, R., De Andres, S.A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics 7*(1) 3.

[35] Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics 20*(18) 3583-3593.