Pertanika J. Sci. & Technol. 26 (1): 329 - 340 (2018)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximising Classification Accuracy

Yap Bee Wah[1]*, Nurain Ibrahim[1], Hamzah Abdul Hamid[1,2], Shuzlina Abdul-Rahman[3] and Simon Fong[4]

[1]Advanced Analytics Engineering Centre, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 UiTM, Shah Alam, Selangor, Malaysia
[2]Institute of Engineering Mathematics, Universiti Malaysia Perlis, Kampus Pauh Putra, 02600 UMP, Arau, Perlis, Malaysia
[3]Centre of Information Systems Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 UiTM, Shah Alam, Selangor, Malaysia
[4]Faculty of Science and Technology, University of Macau, Avenida da Universidade, Taipa, Macau, China

ABSTRACT

Feature selection has been widely applied in many areas such as classification of spam emails, cancer cells, fraudulent claims, credit risk, text categorisation and DNA microarray analysis. Classification involves building predictive models to predict the target variable based on several input variables (features). This study compares filter and wrapper feature selection methods to maximise the classifier accuracy. The logistic regression was used as a classifier while the performance of the feature selection methods was based on the classification accuracy, Akaike information criteria (AIC), Bayesian information criteria (BIC), Area Under Receiver operator curve (AUC), as well as sensitivity and specificity of the classifier. The simulation study involves generating data for continuous features and one binary dependent variable for different sample sizes. The filter methods used are correlation based feature selection and information gain, while the wrapper methods are sequential forward and sequential backward elimination. The simulation was carried out using R, an open-source programming language. Simulation results showed that the wrapper method (sequential forward selection and sequential backward elimination) methods were better than the filter method in selecting the correct features.

Keywords: Feature selection methods, filter method, logistic regression, simulation, wrapper method

ARTICLE INFO

Article history:
Received: 03 March 2017
Accepted: 28 September 2017

E-mail addresses:
beewah@tmsk.uitm.edu.my (Yap Bee Wah),
nurain9270@salam.uitm.edu.my (Nurain Ibrahim),
amz_bst@yahoo.com (Hamzah Abdul Hamid),
shuzlina@tmsk.uitm.edu.my (Shuzlina Abdul-Rahman),
ccfong@umac.mo (Simon Fong)
*Corresponding Author

INTRODUCTION

Classification is one of the most important tasks in many diverse areas such as business, finance, marketing, engineering, medicine,

ISSN: 0128-7680

bio-informatics, and bio-medical engineering. Classification techniques are used to assign subjects to a specific class of a target variable. In classification problems, predictive models are developed to predict the target variable based on several input variables (features). Features, which are also referred to as attributes, are independent variables. Classification problems, such as classification of cancer tumour, images, handwriting, or spam emails, usually involve many features. Therefore, there is continuing research on finding an efficient method to select relevant features with minimal information loss. Classification of data often contains redundant, irrelevant, useless and misleading features. Hence, feature selection plays an important role in solving classification problems (Jirapech-Umpai & Aitken, 2005; Gheyas & Smith, 2010; Yongjun, Minghao, Kiejung, & Keun, 2012; Zhongyi, Yukun, Tao, & Raymond, 2015).

A complex classification problem involves a large number of features. The classifier will take a longer time to classify the observations when the number of features is very large. Several feature selection methods have been developed to solve classification problems. Feature selection methods deal with dimensionality reduction of a large number of features due to irrelevant and redundant features that may negatively affect the accuracy of classification. The main aim of feature selection is to minimise the dimensionality of the features, maximise the accuracy of classification and prevent overfitting.

Most studies have compared feature selection methods using several datasets. This study compares the selected features using a filter and wrapper methods via a simulation study. The selected filter methods are information gain and correlation based feature selection, while the wrapper methods are sequential forward selection and sequential backward elimination. The simulation procedure was carried out using R-an open source programming language. The feature selection methods were then applied to three datasets obtained from the UCI Machine Learning Laboratory.

In Section 2, we discussed some reviews on feature selection methods. The simulation procedure is given in Section 3. The results are presented in Section 4 and Section 5 concludes the paper.

FEATURE SELECTION METHODS

The feature selection methods have a lot of advantages such as reducing the cost of acquiring data and probably making the classification models much easier to understand (Cantú-Paz, 2004). In general, the feature selection methods can be categorised into the filter, wrapper and embedded methods (Ladha & Deepa, 2011; Naqvi, 2012). The advantages and disadvantages of filter, wrapper and embedded methods have been summarised by Ladha and Deepa (2011), Saeys, Inza, and Larranaga (2007), Bolón-Canedo, Sánchez-Maroño, and Alonso-Betanzos (2013), and Bolón-Canedo, Sánchez-Maroño, Alonso-Betanzos, Benítez, and Herrera (2014). Generally, the filter methods are faster and independent of the classifier. Meanwhile, the wrapper and embedded methods are classifier dependent, which means they interact with the classifier. The wrapper methods are simple methods but there may have a risk of overfitting the model. Some examples of the filter methods are chi-square, information gain, correlation based feature selection and relief. The wrapper methods apply searching techniques such as the sequential forward selection, sequential backward elimination, and plus-1-take-away-r with a classifier.

The embedded methods involve classifier such as decision trees, weighted naive Bayes and weight vector of Support Vector Machine (SVM) in Saeys, Inza, and Larranaga (2007), and SVM-RFE (Guyon, Barnhill, & Vapnik, 2002) and kernel-penalized SVM (Maldonado, Weber, & Basak, 2011). Meanwhile, Hui-Huang, Cheng-Wei, and Ming-Da (2011), and Uğuz (2012) and Naqvi (2012) proposed hybrid feature selection by combining the filter and wrapper methods.

The Filter Methods

The filter methods assess the relevance of features by using a ranking procedure that consequently removes low-scoring features. The filter methods are found to be fast, scalable, computationally simple and independent of the classifier. The methods are divided into two categories: the univariate filter method and multivariate filter method. The univariate methods evaluate the features independently, thereby ignoring feature dependencies and leading to poor feature subsets (Yongjun, Minghao, Kiejung, & Keun, 2012; Yusta, 2009). Unlike the univariate methods that ignore feature dependencies and interaction with the classification algorithm, the multivariate methods consider these two factors to a certain degree (Saeys, Inza, & Larranaga, 2007). The first two methods that will be explained in the subsequent section fall into the mutivariate category, while the last two methods fall into the univariate category.

- *Correlation based Feature Selection* - Correlation based feature selection deals with the features that have redundancy among the features. Correlation based feature selection finds features that are highly correlated with the target variable, but have low inter-correlation between the features by using the correlation coefficient (Yongjun, Minghao, Kiejung, & Keun, 2012; Hall, 1999). For correlation based feature selection, the correlation of each pair of features will be calculated. The highest correlation coefficient value will be the first feature to be selected. The equation of correlation based feature selection is (Huiqing, Jinyan, & Limsoon, 2002):

$$M_S = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \quad (1)$$

  where, M_S is the heuristic merit of a feature subset containing k features, $\bar{r}_{cf}$ is the average of the correlation between the features and the target variable, and $\bar{r}_{ff}$ is the average inter-correlation between the features.

- *Fast Correlation-based Filter* - Fast correlation-based filter (FCBF) starts with a full set of features. Fast correlation-based filter uses symmetrical uncertainty to calculate the dependency of features and removes redundant features by using the backward selection method (Zeng, Li, & Chen, 2010). This method has inside stopping criterion to stop it from eliminating the features. Fast correlation-based filter is faster than other feature selection methods. Lei and Huan (2004) provided the algorithm for FCBF method.

- *Information Gain* - A measure based on the information theory of entropy. Entropy is a measure of *disorderliness or noisiness*. Information gain measures the reduction in entropy before and after including the features (Uğuz, 2012; Lei & Huan, 2004). A feature with a high information gain value should be preferred over other features. Information gain does not remove redundant features. The information gain about X provided by Y is calculated as follows:

$$IG(X|Y) = H(X) - H(X|Y) \tag{2}$$

Where,

$$H(X) = -\sum_{i=1}^{k} P(x_i)\log_2 P(x_i) \tag{3}$$

is the entropy of the variable X, and

$$H(X|Y) = -\sum_{i} P(y_j)\sum_{i} P(x_i|y_j)\log_2\left(P(x_i|y_j)\right) \tag{4}$$

is the entropy of X after observing another variable Y. Continuous features need to be discretised when using entropy (Liu, Hussain, Tan, & Dash, 2002).

Each feature will be ranked based on their respective information gain value. Basically, the higher the value, the more informative the feature is.

- *Chi-squared Statistics* – The chi-squared statistics method evaluates association of two categorical variables. Thus, numeric variables need to be discretised into several intervals. The Chi-square statistic is obtained as follows (Huiqing, Jinyan, & Limsoon, 2002):

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\left(\frac{A_{ij} - E_{ij}}{E_{ij}}\right) \tag{5}$$

where m is the number of intervals, k is the number of classes, A_{ij} is the number of samples in the i^{th} interval j^{th} class, R_i is the number of samples in the i^{th} interval, C_j is the number of samples in the j^{th} class, N is the total number of samples, and is the expected frequency of $A_{ij}\left(E_{ij} = R_i * C_j / N\right)$.

Basically, the larger the calculated chi-squared value, the more important the feature is.

The Wrapper Methods

The wrapper methods function almost similar to the filter methods except that they make use of a predefined classification algorithm instead of an independent measure for the subset evaluation. The wrapper methods give a better result compared to the filter methods, but they tend to be more computationally expensive when the number of features becomes very large (Yongjun,

Minghao, Kiejung, & Keun, 2012; Kohavi & John, 1997; Inza, Sierra, Blanco, & Larrañaga, 2002). The first two searching techniques in the wrapper methods described in the subsequent section are the two most common greedy methods frequently employed for feature selection.

- *Sequential Forward Selection* - Sequential forward selection (SFS) starts from the empty set. It performs best when only a small number of features are involved. Nonetheless, the main disadvantage of sequential forward selection is that it is unable to remove features that become insignificant after the addition of other features.

- *Sequential Backward Elimination* - Sequential Backward Elimination (SBE), which is also known as Sequential Backward Selection (SBS), works in the opposite direction of sequential forward selection. Basically, sequential backward elimination starts with a full set of features. Sequential backward elimination works best with a large number of features in the dataset (Ladha & Deepa, 2011).

- *Plus-1-take-away-r* - This method attempts to overcome the nesting effect. In the case of SFS, the nesting effect is a situation whereby once the selected features are selected, they cannot be removed and similar to SBE, once the selected features are removed, they cannot be re-selected. This method allows SFS to use l times forward and then r back-tracking steps of SBS. The challenge with the "plus-l-take-away-r" method is predicting the best (l, r) values to obtain good results with moderate computation (Unler & Murat, 2010).

- *Sequential Floating Forward and Backward Selection (SFFS and SFBS)* - These two methods were introduced by Pudil, Novovičová and Kittler (1994). Backtracking is controlled without any parameter setting. These methods allow a more flexible method since the number of forward and backtracking steps is not predetermined, but instead, it is dynamically changed (Shuzlina, 2012). The SFFS and SFBS are probably the most effective FS methods (as cited in Yusta, 2009). These floating methods allow dynamic addition and deletion of the feature subsets until a suitable number of feature subsets are obtained. The benefit offered by the floating method over the plus-l-take-away-r is its ability to sweep through feature subsets to obtain good results.

SIMULATION PROCEDURES

In this simulation study, data X were simulated and assigned to group 1 or 0 using the following logistic regression model:

$$P(Y=1)=\frac{1}{1+e^{-z}} \tag{6}$$

where $z=\beta_0+\beta_1X_1+\beta_2X_2+\cdots\beta_kX_k$ and k is the number of features. For k=10, we set 6 significant features, while for k=50 we set 20 significant features. The predictors (or features) were set as significant features with odds-ratio greater than 1. For example, the odds-ratios for X5 and X8 are significant features with exp(0.4)=1.087 and exp(0.5)=1.359, respectively.

An odds-ratio close to 1 indicates the feature is not significantly related to Y. Ddata for the logistic regression model were generated using the same technique used by Hosmer and Hjort (2002). The simulation procedure for k=10 is as follows:

1. Generate continuous random features from X1, X2 … to X10 from a standard normal distribution.
2. Calculate z = (0.7 + 0.0000001*x1 + 0.0000001*x2 + 0.0000001*x3 + 0.0000001*x4 + 0.4*x5 + 0.4*x6 + 0.4*x7 + 0.5*x8 + 0.5*x9 + 0.5*x10) and $\pi(x)=\frac{1}{1+e^{-z}}$.
3. Generate the data u from a uniform distribution, U(0,1).
4. Generate outcomes for binary logistic regression by using the rule y=1 if $\mu \leq \pi(x)$ and y=0 otherwise.
5. Apply the feature selection method FSM(j).
6. Count the number of correctly selected features. Repeat 1-6, 1000 times, and obtain an average percentage of the correct features selected.

RESULTS

Results obtained from the simulation study and real datasets are discussed in this section.

Simulation Results

Based on the simulation results in Table 1 and Table 2, the percentages of correctly selected features increase as the sample size increases. The wrapper methods perform better than the filter methods when the model contains 10 features. Meanwhile, the information gain method did not perform well compared to the correlation-based and wrapper methods.

Table 1
Percentage of correctly selected features (k=10, 6 significant features

Feature Selection Methods	Sample Size				
	100	200	300	500	1000
Correlation-based feature selection	74.4%	84.7%	90.7%	96.2%	99.4%
Information Gain	33.3%	33.3%	33.3%	33.3%	35.0%
Sequential Forward Selection	80.3%	89.8%	94.9%	99.2%	100%
Sequential Backward Elimination	80.5%	89.8%	94.9%	99.2%	100%

Table 2
Percentage of correctly selected features (k=50, 20 significant features)

Feature Selection Methods	Sample Size				
	100	200	300	500	1000
Correlation-based feature selection	63.2%	80.5%	88.4%	94.7%	98.7%
Information Gain	0.4%	0.3%	0.4%	0.7%	4.2%
Sequential Forward Selection	58.5%	74.9%	84.4%	93.2%	99.2%
Sequential Backward Elimination	57.0%	75.0%	84.4%	93.2%	99.2%

Application to real datasets

Next, the feature selection methods were applied to three real datasets. The datasets used were Pima Indians Diabetes, Breast Cancer Wisconsin and Spambase obtained from UCI Machine Learning Repository. The sample size for the Pima Indians Diabetes is 768 with eight continuous features. Outcome variable is a binary variable which is denoted as 1 if a patient is tested positive and 0 if it is negative for diabetes. Table 3 summarises the results for the Pima Indians Diabetes dataset.

Table 3
Pima Indians diabetes dataset results

Method/ Performance	AIC	BIC	AUC	ACC	SEN	SPEC
No feature selection (x1*,x2*x3*,x4, x5,x6*, x7*,x8)	741.45	783.24	0.8394	78.2%	58%	89%
Correlation-based feature selection (3 features: x1*, x2*, x6*)	752.12	770.70	0.826	76.69%	57%	87%
Information Gain (2 features: x2*, x6*)	777.4	791.33	0.8109	76.43%	53%	89%
Sequential Forward Selection (6 features: x2*,x6*,x1*,x7*, x3*,x8)	739.46	771.97	0.8348	77.34%	58%	88%
Sequential Backward Elimination (6 features: x1*,x2*,x3*,x6*, x7*, x8)	739.46	771.97	0.8384	77.34%	58%	88%

*Significant feature
Source: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

Based on the results in Table 3, the filter methods selected have fewer significant features compared to the wrapper methods. The correlation based feature selection selected three features (x1, x2, x6), while information gain selected only two features (x2, x6). Meanwhile, both the sequential forward selection and sequential backward elimination selected six significant features. The logistic regression has higher accuracy, sensitivity and specificity with the eight features selected by the wrapper methods.

The sample size for the Breast Cancer Wisconsin Dataset is 669 with nine continuous features. The outcome variable is a binary variable, which is denoted as 1 if the tumour is malignant and 0 if benign. Results for the Breast Cancer Wisconsin Dataset are shown in Table 4. The correlation-based feature selection only selects one variable (x9), while the information gain selected two features (x2 and x7). Meanwhile, both the sequential forward selection and sequential backward elimination selected the first eight features. These results show that wrapper methods, using the sequential forward selection and sequential backward elimination, managed to select more significant features compared to the filter methods. The application to two real datasets confirms the simulation result indicating that the wrapper methods are better than the filter selection methods. The logistic regression has lower accuracy, sensitivity and specificity, with the single feature selected by the correlation based method.

Table 4
Breast cancer Wisconsin dataset results

Method/ Performance	AIC	BIC	AUC	ACC	SEN	SPEC
No feature selection						
(x1*,x2,x3,x4*,x5,x6, x7*, x8, x9)	132.18	177.64	0.9959	97.0%	98.0%	95.0%
Correlation-based feature selection						
(x9*)	735.08	744.19	0.7101	79.0%	97.0%	44.0%
Information Gain						
(x7*, x2*)	232.67	246.32	0.984	93.5%	97.0%	88.0%
Sequential Forward Selection						
(x3,x6*,x1*,x8,x7*,x2, x5,x4*)	134.28	175.23	0.9955	97.0%	98.0%	95.0%
Sequential Backward Elimination						
(x1*,x2,x3, x4*,x5, x6*,x7*,x8)	134.28	175.23	0.9955	97.0%	98.0%	95.0%

*Significant feature
(*Source:* https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes)

The sample size for the Spambase dataset is 4601 with 57 continuous features. Results for the Spambase dataset are shown in Table 5. The outcome is a binary variable which denotes whether the email is considered a spam email (1), or otherwise (0). The result shows that the information gain method selects only two significant features (x52 and x53). Meanwhile, the correlation based method selected sixteen significant features. The sequential backward elimination selected 44 features (3 were not significant, x11, x18 and x22), while the sequential backward elimination selected only ten significant features. The logistic regression performs best with the 44 features selected by the sequential backward elimination with highest accuracy, sensitivity and specificity, lowest AIC and BIC values. This results of this study support the findings by Inza, Larrañaga, Blanco, and Cerrolaza (2002), whereby applications show that the wrapper methods perform better than the filter methods in gene selection in DNA microarray domains.

Table 5
Spambase dataset results

Method/ Performance	AIC	BIC	AUC	ACC	SEN	SPEC
No feature selection						
(x1*, x2*, x5*, x6*, x7*, x8*, x9*, x10*, x12*, x15*, x16*, x17*, x19*, x20*, x21*, x23*, x24*, x25*, x26*, x27*, x28*, x29*, x33*, x35*, x36*, x39*, x41*, x42*, x44*, x45*, x4x6*, x48*, x49*, x52*, x53*, x54*, x5x6*, x57*)	1931.8	2304.94	0.98	93.0%	89.0%	96.0%
Correlation-based feature selection						
(x3*, x5*, x6*, x7*, x8*, x16*, x17*, x18*, x19*, x20*, x21*, x22*, x23*, x24*, x52*, x53*, x57*)	2893.7	3009.54	0.94	89.0%	79.0%	95.0%
Information Gain						
(x52*, x53*)	4472.9	4492.21	0.88	82.0%	61.0%	95.0%
Sequential Forward Selection						
(x51*, x36*, x31*, x15*, x13*, x41*, x29*, x14*, x28*, x32*)	5035.3	5106.1	0.8	73.0%	46.0%	91.0%
Sequential Backward Elimination						
(x1 *, x2* , x3 ,x4 ,x5* ,x6*, x7*, x8*, x9*, x10* , x11 , x12* , x16* , x17* , x18 , x19* , x20* , x21* , x22 , x23* , x24* , x25* , x26* , x27* , x30 , x33* , x34 , x35* , x37 , x38 , x39* , x40 , x42* , x43 , x44* , x45 , x46* , x47 , x48* ,x49* , x50 , x52* , x53* , x54* , x56* , x57*)	1969.8	2272.19	0.98	93.0%	89.0%	96.0%

CONCLUSION

Feature selection methods depend on types of features. This simulation study with continuous features shows that the wrapper method selected more significant features compared to the filter methods. Nonetheless, the information gain did not perform well for continuous features. The application to three datasets confirms that the wrapper method using sequential backward elimination is the best selection method for data with continuous features. The feature selection methods can be easily applied using R, an open source programming language. The simulation study is being extended to compare the performance of the filter and wrapper methods for categorical type features. Future research can also look into feature selection using random forest (Genuer, Poggi, & Tuleau-Malot, 2010), the multivariate-based feature filter method called the kernel PLS-based filter method (Sun, Peng, & Shakoor, 2014) or the hybrid methods Hsu, Hsieh, & Lu, 2011; Uğuz, 2012; Naqvi, 2012; Shilaskar & Ghatol, 2013).

The R-syntax for the simulation study and application of feature selection using R can be obtained from the corresponding author.

REFERENCES

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.

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.

Cantu-Paz, E. (2004). Feature subset selection, class separability, and genetic algorithms. In *Genetic and evolutionary computation conference* (pp. 959-970). Springer Berlin Heidelberg.

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using Random Forests. *Pattern Recognition Letters, 31*(14), 2225-2236.

Gheyas, I. A., & Smith, L. S. (2010). Feature subset selection in large dimensionality domains. *Pattern Recognition, 43*(1), 5–13.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1-3), 389-422.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. (Doctoral Dissertation). The University of Waikato, New Zealand.

Hosmer, D. W., & Hjort, N. L. (2002). Goodness-of-fit processes for logistic regression: simulation results. *Statistics in Medicine, 21*(18), 2723-2738.

Hsu, H. H., Hsieh, C. W., & Lu, M. D. (2011). Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications, 38*(7), 8144-8150.

Huiqing, L., Jinyan, L., & Limsoon, W. (2002). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics, 13*, 51-60.

Inza, I., Larrañaga, P., Blanco, R., & Cerrolaza, A. J. (2002). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine, 31*(2), 91-103.

Inza, I., Sierra, B., Blanco, R., & Larrañaga, P. (2002). Gene selection by sequential search wrapper approaches in microarray cancer class prediction. *Journal of Intelligent and Fuzzy Systems, 12*(1), 25-33.

Jirapech-Umpai, T., & Aitken, S. (2005). Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics, 6*(1), 148-158.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1-2), 273-324.

Ladha, L., & Deepa, T. (2011). Feature Selection Methods and Algorithms. *International Journal on Computer Science and Engineering (IJCSE), 3*(5), 1787 – 1797.

Lei, Y., & Huan, L. (2004). Efficient Feature Selection via Analysis of Relevancy and Redundancy. *Journal of Machine Learning Research, 5*(Oct), 1205-1224.

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery, 6*(4), 393-423.

Maldonado, S., Weber, R., & Basak, J. (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences, 181*(1), 115-128.

Naqvi, G. (2012). *A Hybrid Filter-Wrapper Approach for Feature Selection.* (Master's Thesis). Orebro University, Sweden.

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters, 15*(11), 1119-1125.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507-2517.

Shilaskar, S., & Ghatol, A. (2013). Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. *Expert Systems with Applications, 40*(10), 4146-4153.

Shuzlina, A. R. (2012). *Multivariate filter with particle swarm optimisation variants for feature selection.* (Doctoral dissertation). Universiti Kebangsaan Malaysia, Malaysia.

Sun, S., Peng, Q., & Shakoor, A. (2014). A kernel-based multivariate feature selection method for microarray data classification. *PloS ONE, 9*(7), e102541.

Uğuz, H. (2012). A hybrid system based on information gain and principal component analysis for the classification of transcranial Doppler signals. *Computer Methods and Programs in Biomedicine, 107*(3), 598-609.

Unler, A., & Murat, A. (2010). A discrete particle swarm optimization method for feature selection in binary classification problems. *European Journal of Operational Research, 206*(3), 528-539.

Yongjun, P., Minghao, P., Kiejung, P., & Keun, H. Y. (2012). An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. *Bioinformatics, 28*(24), 3306-3315.

Yusta, S. C. (2009). Different metaheuristic strategies to solve the feature selection problem. *Pattern Recognition Letters, 30*(5), 525-534.

Zeng, X. Q., Li, G. Z., & Chen, S. F. (2010, December). Gene selection by using an improved Fast Correlation-Based Filter. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on* (pp. 625-630). IEEE.

Zhongyi, H., Yukun, B., Tao, X., & Raymond, C. (2015). Hybrid filter–wrapper feature selection for short-term load forecasting. *Engineering Applications of Artificial Intelligence, 40*, 17–27.