An Analysis of Hybrid Feature Selection Algorithms' Performance in Classification

Meng Zhang
University of Toronto
Toronto, Canada
mzhan63@cs.toronto.edu

Abstract

Using classification models to predict or classify real-world problems is becoming more and more popular. But the high dimensionality can increase the memory storage cost, computation complexity and affect the performance of a classification model. Tremendous effort has been devoted to exploring methods to reduce dimensions. One of the methods is through feature selection. In this paper, nine datasets with different characteristics are examined on different hybrid and basic feature selection algorithms in both linear and non-linear classification models. The experiment result indicates that hybrid feature selection algorithms have some apparent advantages over using basic algorithms alone but also shows nonnegligible concerns in using them. Suggestions in using hybrid feature selection methods are given in conclusion section.

1 Introduction

Machine learning aims to automate the process of information discovery and be of use to people from a wide range of backgrounds without the requirement of domain knowledge [Hall and Smith, 1997]. The efficiency of finding the most suitable model to a represent a dataset not only relies on the machine learning model but also counts on how predictive the features are. However, a large amount of data comes with a lot problem. Dimensionality is one of the biggest problems.

Dimensionality constitutes a serious obstacle to the efficiency of most machine learning algorithms. This obstacle is also known as the "curse of dimensionality" [Chizi and Maimon, 2009]. Under certain conditions, the presence of redundant and irrelevant information may result in slowed execution, less understandable results and much-reduced accuracy [Hall and Smith, 1997].

Single feature selection algorithms have been developed for a long time, such as filters, wrappers, and embedded methods. These methods have different advantages over feature selection as well as disadvantages. It is natural that people started wondering what is the effect if we combine them. Thus, the hybrid method becomes very attractive in reducing the drawbacks and joins the strengths of different selection algorithms [NAQVI, 2011].

Considered the challenges, benefits, and questions arose from the hybrid feature selection method; I conducted an empirical experiment to analyze the performance of hybrid feature selection methods in classification problems.

Submitted to 29th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

2 Literature Review

2.1 Fundamental Work

Blum and Langley [1997] discussed the problem of irrelevant features and irrelevant examples. It points out the importance to decide which features to use in describing the concept as one of the major tasks in machine learning. They stated the biggest advantage and disadvantage of the filter as efficient but less accurate. And wrapper's pros and cons are the opposite of filter's.

In the same year, Kohavi and John [1997] compared wrapper approach to filter approaches and explored the relation between optimal feature subset selection and relevance in their work. It identifies the problem of using filter method to define relevance independently of the learning algorithm, points out that wrapper approach requires a search space, operators, a search engine and an evaluation function which make it may only be useful to a specific learning algorithm and shows the overfitting problem caused by using wrapper method.

Guyon and Elisseeff [2003]'s work gives a thorough introduction to variable and feature selection methods. It discussed the limitations and applications of each feature selection methods. This paper also gives some guidance on designing my experiment.

2.2 Related Work

The work of NAQVI [2011] is the most related one to my experiment with the difference that he used QPFS filter method and Sequential Feature Selection wrapper in his experiment. His findings show that feature selection for supervised machine learning can be achieved by utilizing the efficiency of filters and the accuracy of wrappers.

Uncu and Türkşen [2007] proposed a novel feature selection approach in their paper which blends wrapper and filter concept with KNN model. It is not quite the same idea as hybrid methods which will be discussed in my experiment; their approach only includes either a wrapper or a filter method. But their work shares some similarities with mine. Their works' limitation is obvious such as the method is only verified with KNN model. They declared by blending feature wrapper and filter can avoid overfitting problem.

The work by Janecek et al. [2008] investigated the relationship between various feature reduction methods (feature subset selection and dimensionality reduction) and the resulting classification performance. They found the classification accuracy is highly sensitive to the type of data, while the classification accuracy achieved with reduced feature sets is often better than with the full feature set.

3 Problem Definition

Feature subset selection: given a feature set $Y = \{y_1, y_2, ..., y_d\}$, find a subset X_k, with k<d, that maximizes an objective function $J(X_k)$.

$$X_k = \{x_1, x_2, ..., x_k\} = arg \max_k J\{x_j | j = 1, 2, ..., k, ; x_j \in Y\} \tag{1}$$

4 Methods and Models

4.1 Basic Feature Selection Methods

The feature selection algorithms can be divided into three main categories: Wrappers, filters and embedded methods. In this experiment, only selected wrappers and filters are considered. They will be referred by the abbreviations in brackets in the rest sections of this report.

4.1.1 Filter Method

Pearson's Correlation Coefficient (Corr) The Pearson's Correlation Coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. It measures the linear relationship between two variables [9, 2017]. In this experiment, I use the absolute value of the correlation coefficient because of the higher the absolute value, the more correlated the feature and the target. For datasets X and Y, the Pearson's Correlation Coefficient is calculated as follow:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X, \sigma_Y} \tag{2}$$

Mutual Information (MI) Normalized Mutual information is used to measure the non-linear dependence between two variables. The mutual information of dataset X given Y means the reduction of uncertainty of X due to Y [Latham and Roudi, 2009]. The higher the mutual information gain value, the more dependent X and Y are. The calculation of the mutual information between datasets X and Y is as below:

$$MI(X,Y) = \sum_{i=1}^{|} X| \sum_{j=1}^{|} Y| \frac{|X_i \cap Y_j|}{N} \log \frac{N|X_i \cap Y_j|}{|X_i||Y_j|} \tag{3}$$

$|X_i|$ is the number of samples in X_i, $|Y_i|$ is the number of samples in Y_i.

4.1.2 Wrapper Method

Sequential Backward Elimination (SBE) Sequential Backward Elimination starts with the set of all variables and progressively eliminates the least promising ones [Pudil et al., 1994].

Sequential Backward Floating Selection (SBFS) Sequential Backward Floating Selection can be viewed as an extension to SBE algorithm. The floating algorithm has an additional inclusion step to include a feature maybe exclude in previous rounds. Such a feature is only included when it produces a better result when adding to current subset [Pudil et al., 1994].

The algorithm of SBFS method is shown as in Algorithm 1. SBE follows similar algorithm without Step 2. It is expected that SBFS should perform better than SBE in classification accuracy but takes much longer computation time.

Input: the set of all features,Y = $\{y_1, y_2, ...y_d\}$, criterion function $J(X_k)$, preset boundary for size of feature subset.
Output: $X_k = \{x_j | j = 1, 2, ..., k; x_j \in Y\}, where\ k = (1, 2..., d), k < d$

initialization: $X_0 = Y, k = d$;
Step 1 (Exclusion):
$x^- = argmaxJ(x_k - x), wherex \in X_k$
$X_k - 1 = X_k - x^-$
k=k-1
go to Step 2;
Step 2 (Conditional Inclusion):
$x^+ = argmaxJ(x_k + x)$, where $x \in Y - X_k$
if $J(x_k + x) > J(x_k)$ **then**
| $X_{k+1} = X_k + x^+, k = k + 1$
end
go to Step 1;
Termination: the size of feature subset has reached the preset boundary. ;

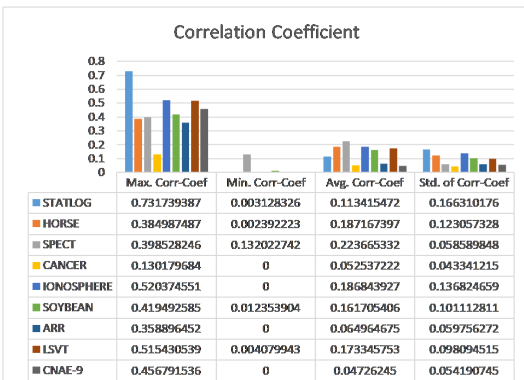
Algorithm 1: Sequential Backward Floating Selection (SBFS)

4.2 Hybrid Method

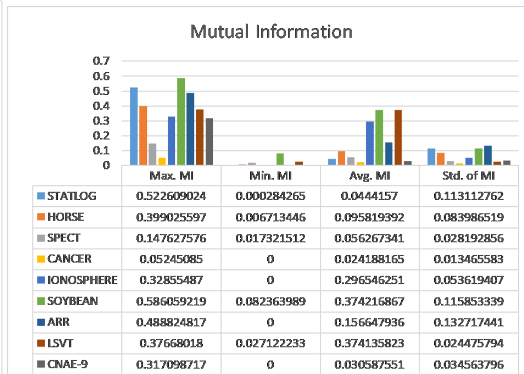
The hybrid feature selection method refers to combining two or more feature selection algorithms to produce the feature subset [NAQVI, 2011]. The algorithm order in this experiment is applying filter method to full data first to produce a subset, A. Then applying wrapper method to A to produce a subset B to be used in model fitting. Since there are two filter methods and two wrapper methods used, it results in four hybrid methods in total: Corr + SBE, Corr + SBFS, MI + SBE, and MI + SBFS.

Dataset	instances amount	features amount	Feature/ Instance ratio	Categorical feature ratio	Real value feature ratio	missing value percent	classes
STATLOG	1000	20	0.02	0.65	0.35	0	2
HORSE	366	22	0.06	0.64	0.36	0.226395	3
SPECT	267	22	0.08	1.00	0.00	0	2
CANCER	858	32	0.04	0.00	1.00	0.1277737	2
IONOSPHER	351	34	0.10	0.00	1.00	0	2
SOYBEAN	307	35	0.11	1.00	0.00	0.095	19
ARR	443	279	0.63	0.26	0.74	0.0032819	10
LSVT	126	310	2.46	0.00	1.00	0	2
CNAE-9	1080	856	0.79	0.00	1.00	0	9

Table 1: Basic Characteristics



(a) Absolute Pearson Correlation Coefficient Statistics for each dataset.



(b) Mutual Information Statistics for each dataset.

Figure 1: Corr and MI Distribution

4.3 Classification Model

Support Vector Machine is used in this experiment because of its ability to deal with both linear separable problems and non-linear separable ones [Gunn et al., 1998]. Two kernels, linear and RBF are used. C (soft margin hyperparameter) is tuned. Although gamma value is tunable for RBF kernel, considering the different number of features of datasets, the gamma value is set to one divided by the number of features in the dataset.

5 Datasets

5.1 Dataset Characteristics

Nine datasets from UCI [13, 2017] are used in this experiment. Their description can be found in Appendix A. I select datasets from different domains to avoid bias and try to cover different variations in feature/instance ratio, categorical features ratio; real value features ratio, missing value ratio, number of classes, correlation coefficient distribution and mutual information distribution. Details can be found in Table 1 and Figure 1. Datasets will be referred by their abbreviations shown in Table 1 in the rest of this report.

5.2 Pre-processing

The original datasets providers convert all categorical feature value in datasets. Label Encoding is applied to SOYBEAN, HORSE, and STATLOG. One Hot Encoding is applied to SPECT and ARR.

Only the classes of IONOSPHERE are characters ‘g’ for good and ‘b’ for bad in original data. I convert these two categories to 1 for good and -1 for bad. ARR originally has 16 classes, but 6 of these classes have less than five instances in the dataset. To use k-fold cross-validation the least occurrence of a class must be greater or equal than the k value. Thus these six classes are removed from ARR.

SOYBEAN and ARR’s missing value is filled with ‘?’ in original data. I impute all these missing data with the mean value of the column (feature).

6 Experimental Design

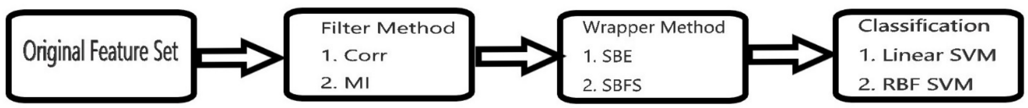


Figure 2: The stages of the hybrid feature selection method and available options in each step.

I used the classification model without any feature selection process as the baseline model for each dataset.

The datasets are divided into training set (80%) and test set (20%). The test set is only used to evaluate the models which are selected after cross-validation on hyperparameters. All classification models are cross-validated in 5-fold and tuned for C value from 0.01, 0.1, 1, -0.1 and 0.1.

The accuracy rate is simply measured by dividing the number of correct predictions by number of test cases.

A total of eight scenarios could happen to each dataset (see Appendix B for a detailed list of all eight scenarios). Due to the distinctions of datasets, some datasets did not complete all scenarios. The experiment workflow and options available in each step is shown as in Figure 2. For each scenario, I collect performance data of the hybrid selection method, the single basic selection method (which is used in the hybrid method) and the baseline model’s performance.

7 Experiment Result and Analysis

Because of the length limitation of this report, all experimental data diagrams are included in the Appendix C at the end of this report. Duplicate cases (thresholds in a scenario return the same feature subset after filter stage) are omitted in the final result. Source code, raw input data, raw experimental data, organized experimental data can all be found in my Github repository for this project [Zhang, 2017]. Please refer to these supplementary materials for the detailed experiment result. A summary of the experiment result is discussed here.

7.1 Test Accuracy

To compare how much the hybrid method can improve the testing performance given all other settings the same, I counted the rate when hybrid method’s accuracy is better than or equal to the baseline model, the filter only model and the wrapper only model. The summary is listed in Table 2. The expectation is in most cases the hybrid method is supposed to beat baseline model and filter model in test accuracy due to the advantage of wrapper method in taking more possible subsets of features into consideration. But the hybrid method may lose to applying wrapper method only because the latter one has more features as input. The common characteristic of all cases where hybrid method beats the other models are when the threshold is low which indicates an adequate number of relative features remain after filter stage.

A special case of the dataset CANCER draws my attention. The hybrid method does not improve its test accuracy compared to any other three models. By comparing with other datasets, I conclude that this is caused by the low average and standard deviation of correlation coefficient and mutual information of CANCER’s features. Both of them are the lowest across all nine datasets. In this case,

Dataset	Test Accuracy Improvement			Same Test Accuracy		
	Hybrid vs. Baseline	Hybrid vs. Filter	Hybrid vs. Wrapper	Hybrid vs. Baseline	Hybrid vs. Filter	Hybrid vs. Wrapper
STATLOG	80%	30%	50%	0%	70%	0%
HORSE	40%	60%	0%	40%	40%	10%
SPECT	50%	42%	67%	0%	42%	17%
CANCER	0%	0%	0%	100%	100%	50%
IONOSPHERE	11%	29%	11%	21%	43%	32%
SOYBEAN	3%	11%	6%	11%	56%	25%
ARR	64%	50%	57%	29%	43%	14%
LSVT	71%	57%	64%	14%	29%	14%
CNAE-9	0%	11%	N/A	0%	50%	N/A

Table 2: Test Accuracy Summary

Dataset	Average Train and Test Accuracy Gap (Train Accuracy - Test Accuracy)			
	Hybrid	Baseline	Filter	Wrapper
STATLOG	-0.0075	0.136875	0.006	0.013125
HORSE	0.268770826	0.385042577	0.358145131	0.283043317
SPECT	0.05272996	0.151538863	0.120131282	0.193075117
CANCER	0.019703515	0.01574683	0.01574683	0.023043935
IONOSPHERE	0.043369143	0.087676056	0.047905289	0.077804326
SOYBEAN	0.003603959	0.016309724	0.001939936	0.044624475
ARR	0.067624307	0.04529296	0.058251943	0.04529296
LSVT	0.096758242	-0.129230769	-0.121208791	0.25
CNAE-9	0.033771495	0.039351852	0.029100529	N/A

Table 3: Train Accuracy Summary

the filter is not very effective in screening out unrelated variables. The highest test accuracy results from only applying SBFS to the dataset includes 5 out of the 32 features. All other cases have the same test accuracy.

A similar situation happens to CNAE-9 as well but for a different reason. The hybrid method does not beat baseline model and only beats filter method in 11% cases. This result is due to the domain of CNAE-9 is text. Thus, all of its features are word frequencies, and 99.22% values are 0. For datasets with these characteristics, using all available features together to predict the class is more powerful than using subsets.

7.2 Train Accuracy

I compared the average gap between train accuracy and test accuracy for each dataset using different algorithm selection methods as shown in Table 3. As expected, hybrid method reduced the overfitting caused by wrapper method in most cases.

The only exception happens in ARR, where hybrid method finds the optimal subset contains only one or two features. It is a side effect caused by filter method's feeding wrapper method in a hybrid method. The filter stage breaks some potential good combinations of features. Thus, the wrapper stage has the chance to find a worse subset than using wrapper method alone and introduced the possibility to enlarge the predictive power of some features which can mislead wrapper stage. In addition, ARR has a very uneven distribution over the ten classes. One of them occupied more than half of the instances. This may cause the other classes' instances can not get trained enough. Specifically, the largest difference in gap is 0.116 for hybrid and 0.045 for the wrapper. This one special case alerts us that hybrid not always helps avoid overfitting. More attention should be paid when the filter method only leaves few features for wrapper method. Although out of all nine datasets, ARR is the only one has this problem, we still need to be alert to such situations when using a hybrid method.

7.3 Execution Time

The execution time follows a general trend that wrapper > hybrid> filter. Execution time for baseline model lies between the fastest to between hybrid and filter for most cases. There are also some cases where the hybrid method takes less time than baseline model if filter stage screen out a large proportion of features (usually over 86.4%) in some cases. Within these cases, hybrid method generate better test performance in more than half of them. Hybrid method execution time is largely affected by how many features remained after filter stage, thus affect how much faster it is than applying wrapper method only. This is the same as my expectation.

7.4 SBE vs. SBFS in Hybrid Methods

It is expected from their algorithms that given all the other conditions the same, SBFS should always perform better or equal fitting to training data and takes longer time than SBE when using alone or included in the hybrid method. The size of optimal feature subset found by these two algorithms does not have a definite quantitative relationship. Because SBFS goes through more feature combinations, and the best one could be greater, equal or even smaller than the subset found by SBE. The experimental data supports this point. By comparing the optimal subset size between hybrid methods use SBE or SBFS, I found in 18 out of 94 comparable cases, hybrid methods use SBFS produce smaller subset; equal size in 66 of the comparable cases; in the rest 10 cases, SBFS hybrid methods result in larger variable size. Although SBFS has its advantage in producing a better subset of features, it does not guarantee smaller subset. In another word, SBFS may not help in reducing the data size needed for a machine learning problem.

Even so, the result in ARR is very impressive. The scenario MI+SBFS+RBF SVM at threshold equals 0.1 has 12 features compared to 162 features when substitutes SBFS with SBE. With much fewer features, the hybrid method with SBFS produces 17% higher test accuracy than the hybrid method with SBE. The execution time of SBFS is around 50% longer than SBE in this case. ARR dataset aims at distinguishing between the presence and absence of cardiac arrhythmia and classify it into different groups. If SBFS method gives a more meaningful subset of features, it may help to identify the key features of each cardiac arrhythmia group, contributes to quicker classification and fewer data required from patients. If time permitted, SBFS is very desirable in looking for the best subset of features.

7.5 Dataset Characteristics

The comparison between datasets reveals some implications. These implications could be occasional and random, so I did a second run on some datasets to verify them.

Feature/instances Ratio: I compared IONOSPHERE, LSVT and CNAE-9, these three datasets have all real value features, no missing value but different feature/instance ratio at 0.1, 2.46 and 0.79. I expected the lower the ratio, the better the test performance because more instances could be used to cross-validate the best subsets. But the result of these three datasets is different with my expectation. It appears like the ratio does not affect feature selection inside the dataset. It is probably because all features are still competing under the same environment. Thus a fair result is still maintained.

Missing Value: STATLOG and HORSE have similar characteristics, but the HORSE has a missing value rate of 22.6% and smaller improvement by hybrid method compared to STATLOG. The same as SPECT compared to SOYBEAN. Such result is reasonable because of the more complete data, the more suitable fitting. Hybrid method consumes fewer features in wrapper stage thus it is more important to have accurate data values.

Categorical and Real Value Features Ratio: My expectation is real value features, and one hot-encoding categorial features should be more accurate than label-encoding categorical features because label encoding usually misleads the importance of these encoded features, i.e., the label assigned to each category does not represent their value in the model. The ratio of different types of features does not show clear pattern in affecting the performance of hybrid method in this experiment.

Certain datasets do not favor feature selection such as CANCER and CNAE-9 mentioned in test accuracy section. Their features are almost equally relative to the target variable. Also, CANCER has the missing rate at 12.78%, CNAE-9 has the sparse rate at 99.22%. Such datasets have better accuracy when using the full set of available features.

8 Limitation

It appeared in the experiment process that when absolute correlation coefficient value has a low variance or low average in a dataset, the filter stage is not very effective in reducing the size of the feature subset. In such case, the hybrid method performs poorly in improving execution time and accuracy compared to other situations. The same problem happens to mutual information filter too.

Sequential selection methods take a long time in execution thus some datasets do not have complete data available for using wrapper method alone.

The effect of categorical features encoding method is not clear in this experiment; more datasets focus on distinguishing this factor is needed to make a more definite conclusion.

The overfitting happens partially due to the exhaustive approach by SBE and SBFS algorithms but may also result from the tuning process on hyperparameters.

Some datasets have very low accuracy in every scenario; the unfit classification models might cause this.

9 Future Work

The threshold of filter stage could be dynamically calculated based the distribution of correlation coefficient value or mutual information score, as well as the number of original features to produce a more manageable and meaningful subset of features for wrapper stage.

In my next study, processing redundant variables in filter stage will be added to help to reduce the size of the subset of features feeding to wrapper stage and improving the computational time. Methods may include keep the best variable in a group of redundant variables or group them and assign orders of trying in wrapper stage.

Experiment with more datasets with different categorical encoding methods may make the effect of encoding clearer. More hyperparameter value could be cross-validated in a future experiment.

10 Conclusion

The hybrid method has shown its advantages in reducing overfitting, reducing execution time, and increasing classification accuracy than using filter or wrapper method alone. Nevertheless, the weaknesses and exceptions of hybrid methods cannot be ignored. Suggestions when using hybrid feature selection methods based on the result of this experiment are listed below in order to deal with these weaknesses as well as taking advantage of its strengths. Some of them may apply to other feature selection methods too.

1. Analyze the distribution of values used in filter stage before setting threshold for hybrid method to keep an adequate number of features for wrapper stage;

2. Use more effective cross-validation hyperparameter values;

3. If possible, use domain knowledge or ask expert in the domain to help identify suitable ways to clean, impute, encode data before processing;

4. For datasets with a large number of features such as ARR, LSVT and CNAE-9 have hundreds or more features, using hybrid method compared to the baseline model to find the optimal subset can save a lot of time and keep a competitive performance compared to using wrapper method alone.

5. Be aware that not all datasets are suitable for feature selections. Datasets with a low variance of features' correlation to target variable or with high missing value rate/sparse value rate, should be treated with more conscious.

References

Uci machine learning repository, 2017. URL http://archive.ics.uci.edu/ml/datasets.html.

Pearson correlation coefficient — wikipedia, the free encyclopedia, 2017. URL https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&oldid=812519727. [Online; accessed 18-December-2017].

A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1):245–271, 1997.

B. Chizi and O. Maimon. Dimension reduction and feature selection. In *Data mining and knowledge discovery handbook*, pages 83–100. Springer, 2009.

S. R. Gunn et al. Support vector machines for classification and regression. *ISIS technical report*, 14: 85–86, 1998.

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

M. A. Hall and L. A. Smith. Feature subset selection: a correlation based filter approach. 1997.

A. Janecek, W. Gansterer, M. Demel, and G. Ecker. On the relationship between feature selection and classification accuracy. In *New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, pages 90–105, 2008.

R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2): 273–324, 1997.

P. E. Latham and Y. Roudi. Mutual information, 2009. URL http://www.scholarpedia.org/article/Mutual_information.

S. NAQVI. A hybrid filter-wrapper approach for featureselection, 2011.

P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125, 1994.

Ö. Uncu and I. Türkşen. A novel feature selection approach: combining feature wrappers and filters. *Information Sciences*, 177(2):449–466, 2007.

M. Zhang. Github repository for materials of this project, 2017. URL https://github.com/zhan4806/CSC2515.

Appendices

A Datasets Description

German Credit Data (STATLOG): This dataset classifies people described by a set of attributes as good or bad credit risks.

Horse Colic (HORSE): This dataset aims at predicting whether a horse can survive based on past medical conditions. It has 22 attributes and six target variables. I only used the outcome target variable in this experiment.

SPECT Heart (SPECT): This dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images.

Cervical cancer (CANCER): This dataset focuses on the prediction of indicators/diagnosis of cervical cancer. The features cover demographic information, habits, and historic medical records.

Ionosphere (IONOSPHER): The task of this dataset is to classify if a given radar signal targets a "good" or "bad" electron.

Soybean (SOYBEAN): Each instance describes properties of a crop of soybeans, and the task is to predict which of the 19 diseases the crop suffers.

Arrhythmia (ARR): The task of this dataset is to distinguish between the presence and absence of cardiac arrhythmia and classify it in one of the 16 groups.

LSVT Voice Rehabilitation (LSVT): This dataset aims at assessing whether voice rehabilitation treatment leads to phonations considered "acceptable" or "unacceptable."

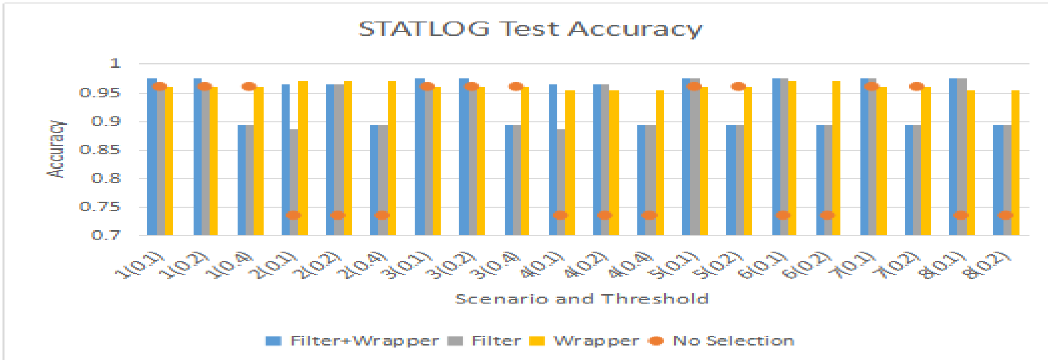
CNAE-9 (CNAE): This is a data set containing 1080 documents of free text business descriptions of Brazilian companies categorized into a subset of 9 categories. CNAE has a high sparse rate at 99.22% due to its text domain characteristic. All these values are filled with 0 in original data.

B Hybrid Scenario Numbering and Details

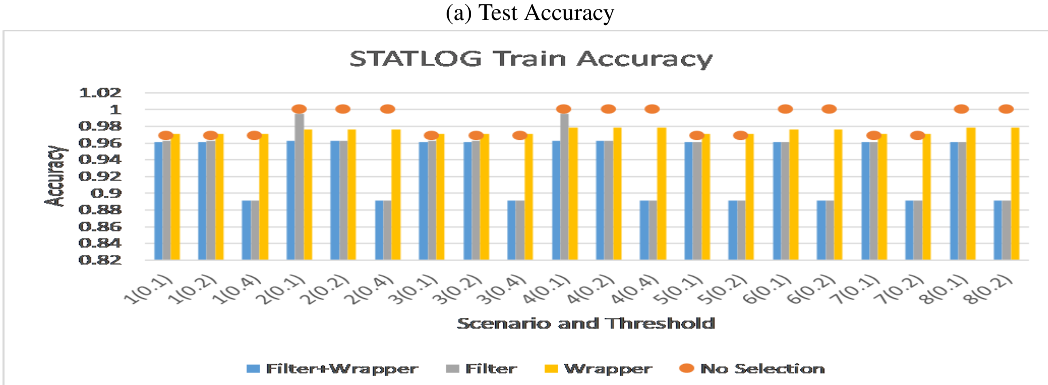
Scenario ID	Filter Method	Wrapper Method	Classification Model
1	Corr-coef	SBE	SVM Linear
2	Corr-coef	SBE	SVM Non-linear
3	Corr-coef	SBFS	SVM Linear
4	Corr-coef	SBFS	SVM Non-linear
5	MI	SBE	SVM Linear
6	MI	SBE	SVM Non-linear
7	MI	SBFS	SVM Linear
8	MI	SBFS	SVM Non-linear

C Experiment Result for Each Dataset

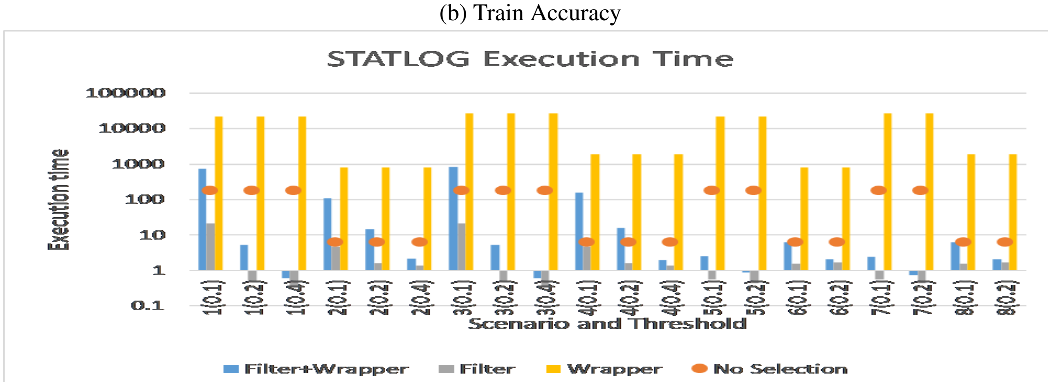
Labels on x-axis are scenario ID and filter threshold (in brackets).



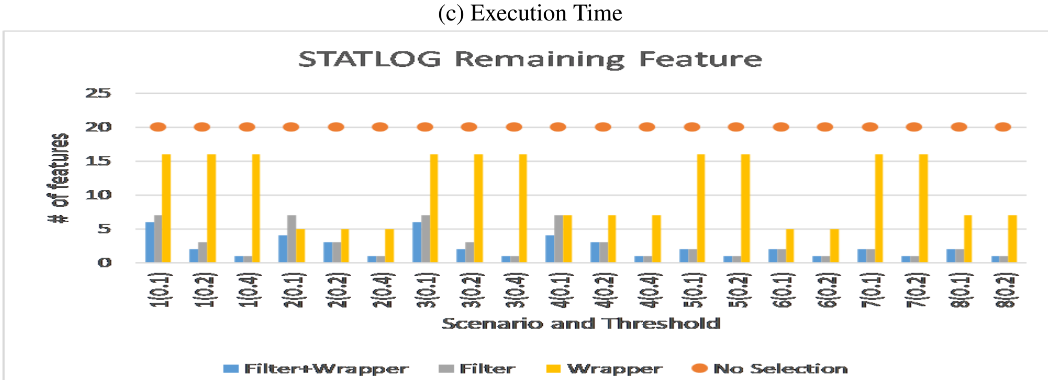
(a) Test Accuracy



(b) Train Accuracy

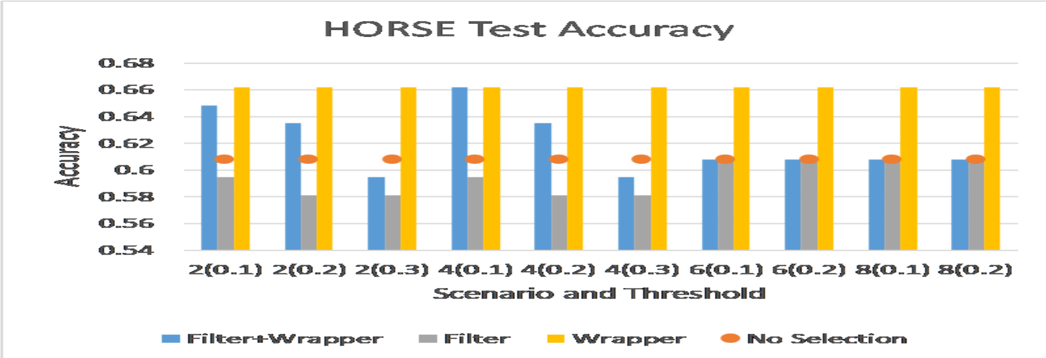


(c) Execution Time

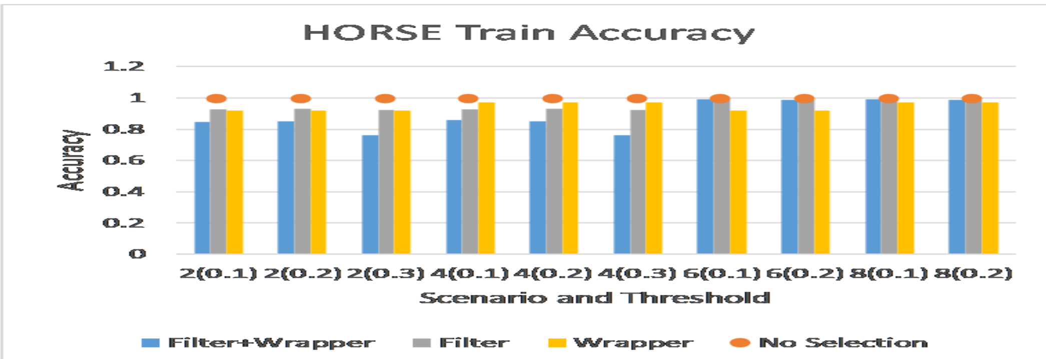


(d) Feature Subset Size

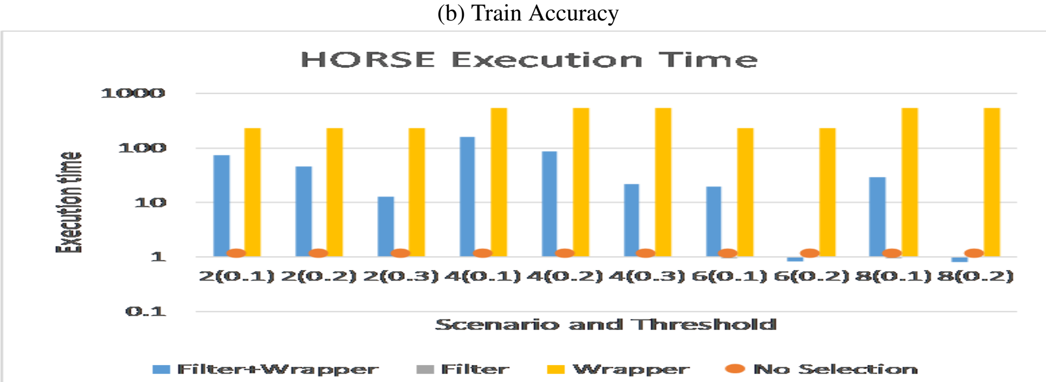
Figure 3: STATLOG Experiment Data



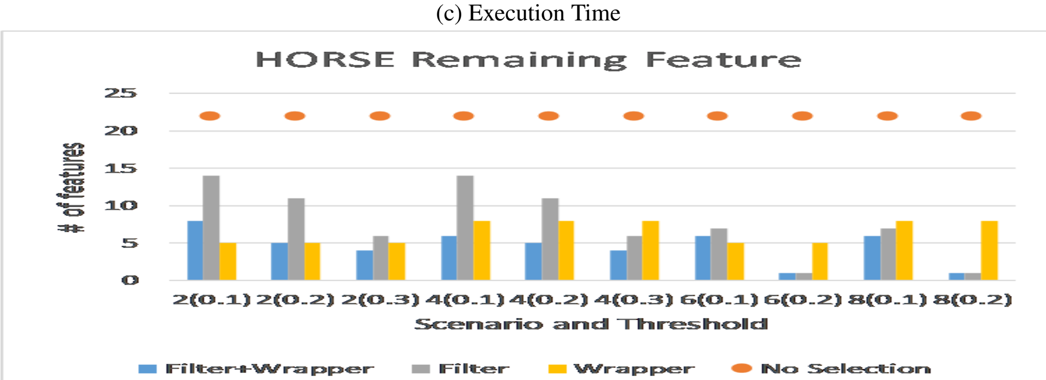
(a) Test Accuracy



(b) Train Accuracy

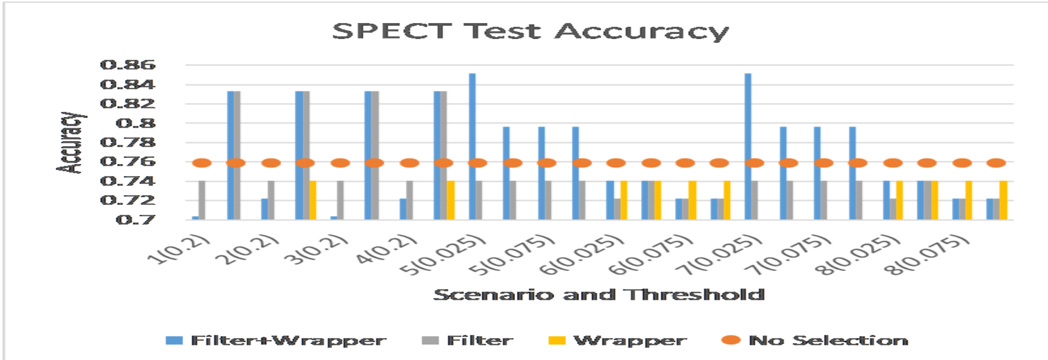


(c) Execution Time

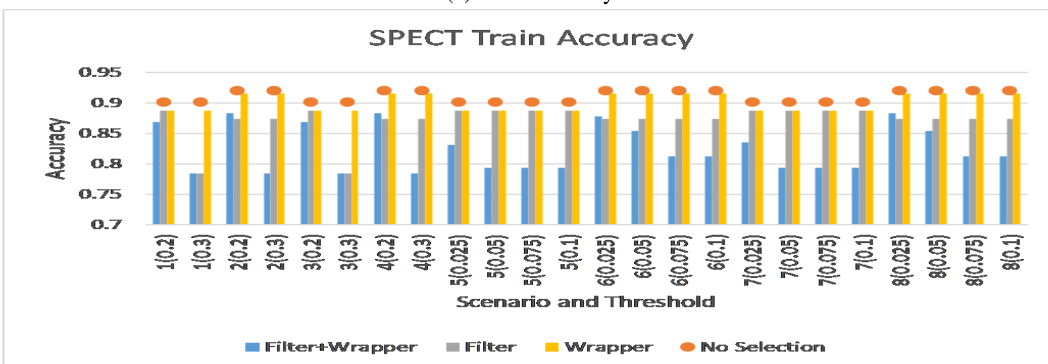


(d) Feature Subset Size

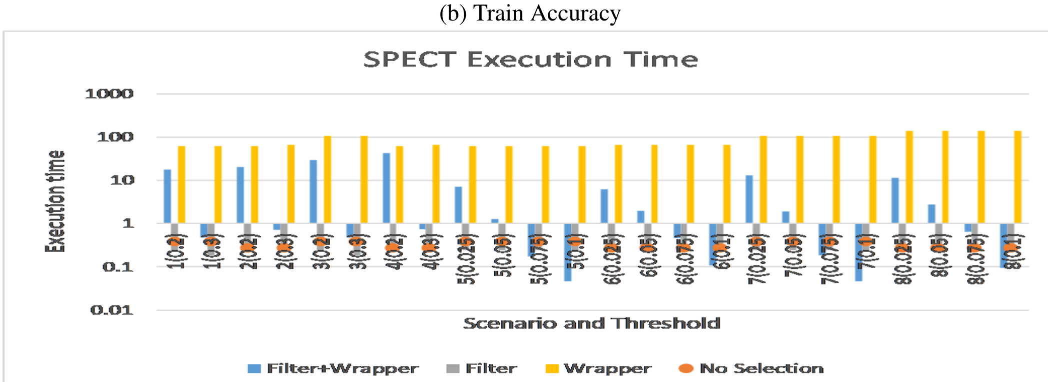
Figure 4: HORSE Experiment Data



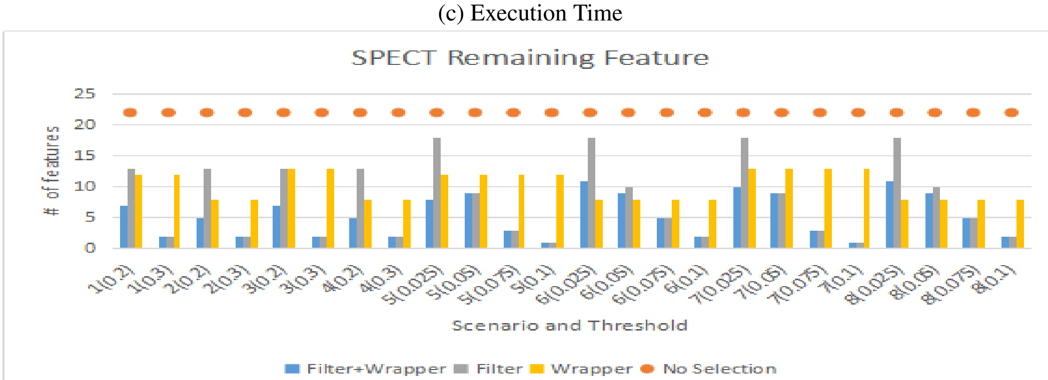
(a) Test Accuracy



(b) Train Accuracy

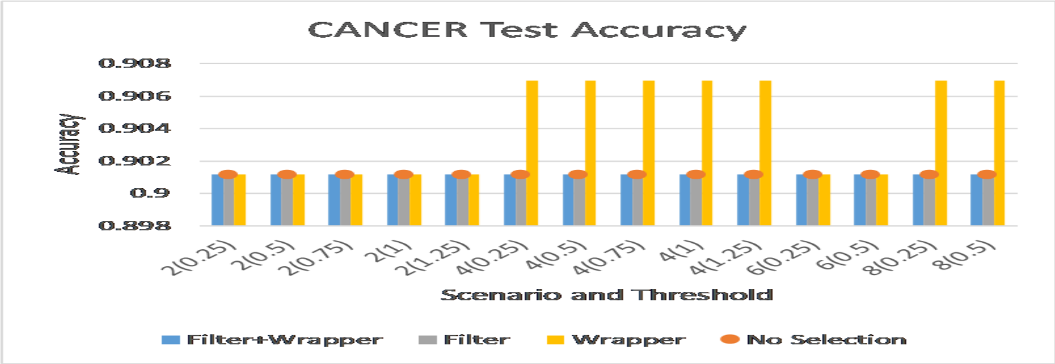


(c) Execution Time

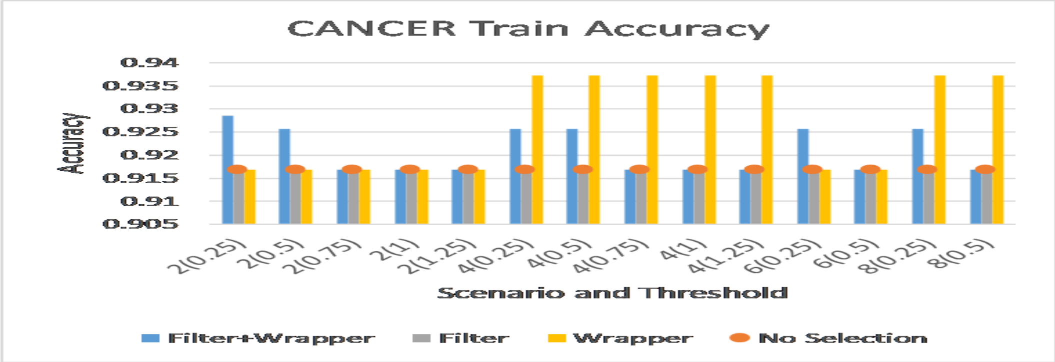


(d) Feature Subset Size

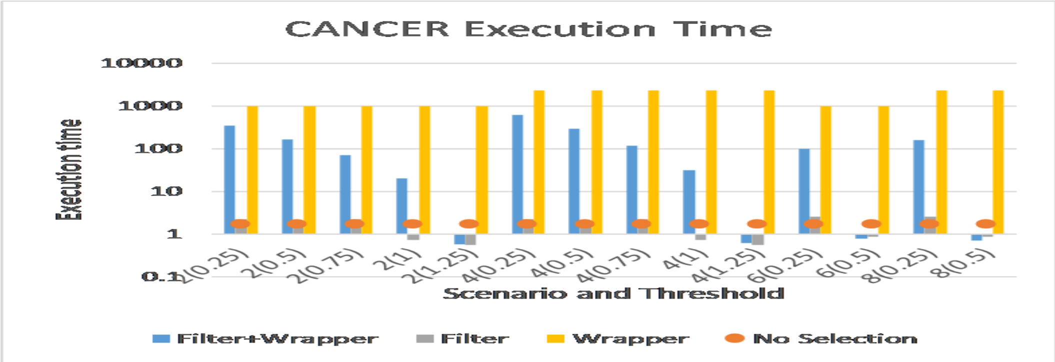
Figure 5: SPECT Experiment Data



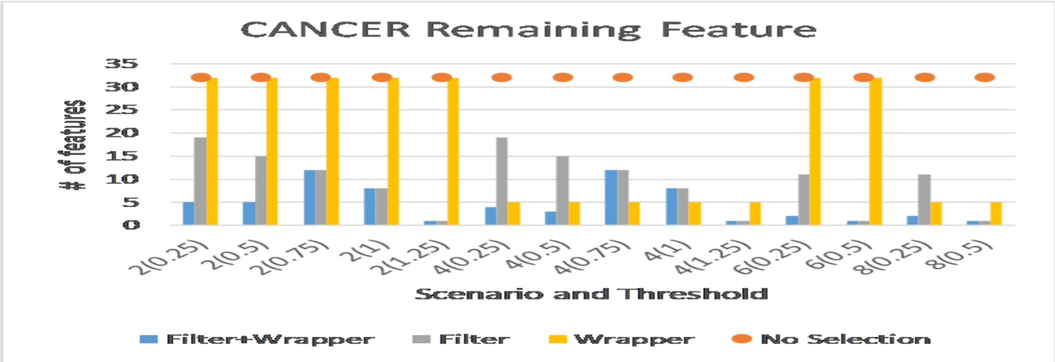
(a) Test Accuracy



(b) Train Accuracy

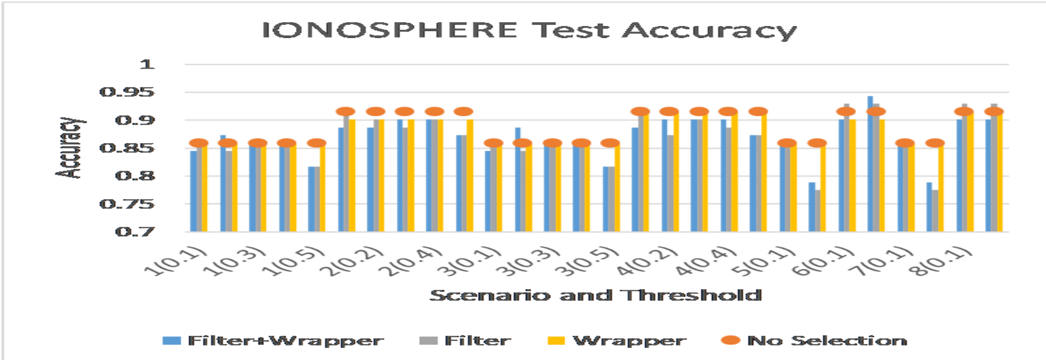


(c) Execution Time

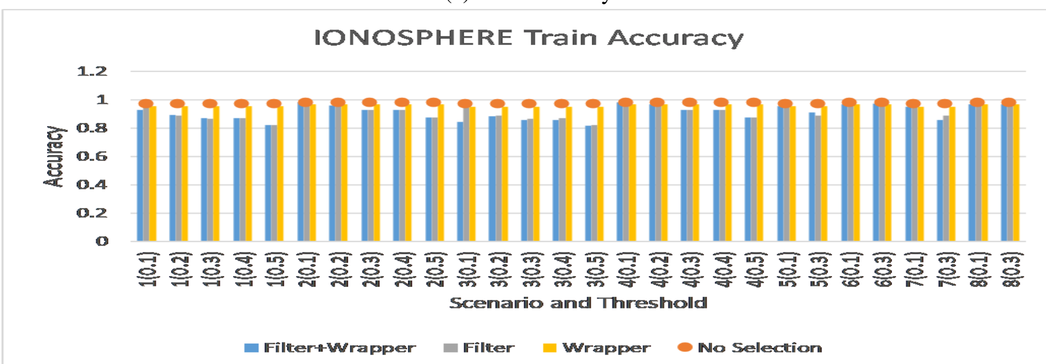


(d) Feature Subset Size

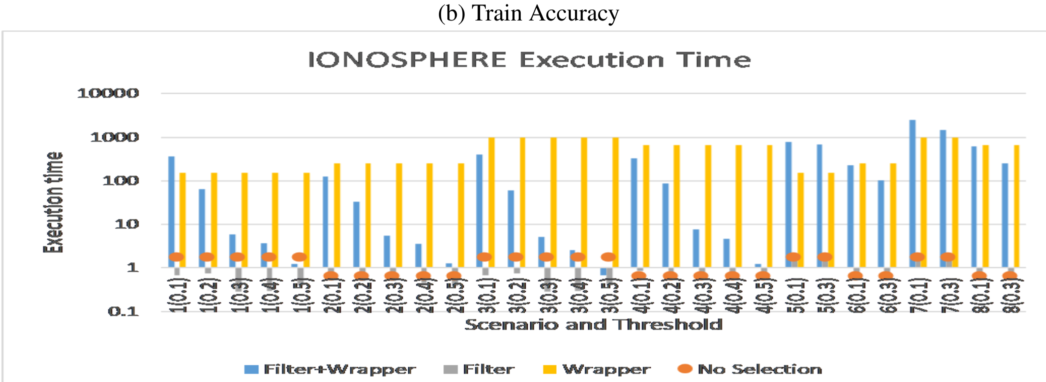
Figure 6: CANCER Experiment Data



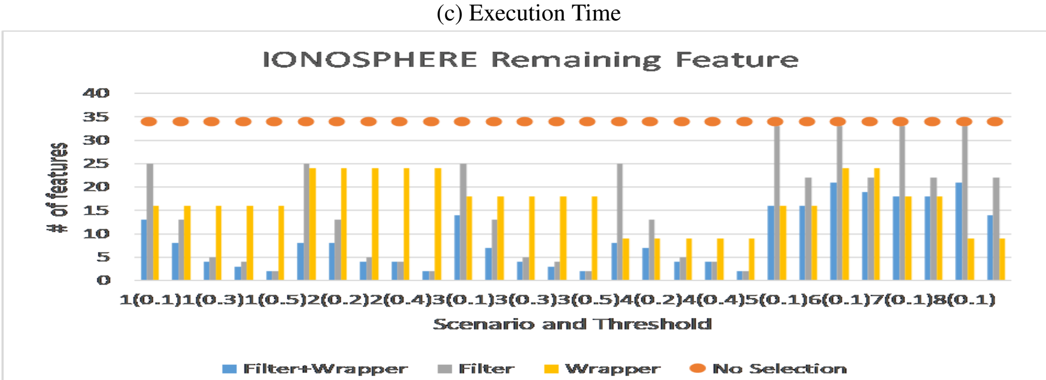
(a) Test Accuracy



(b) Train Accuracy

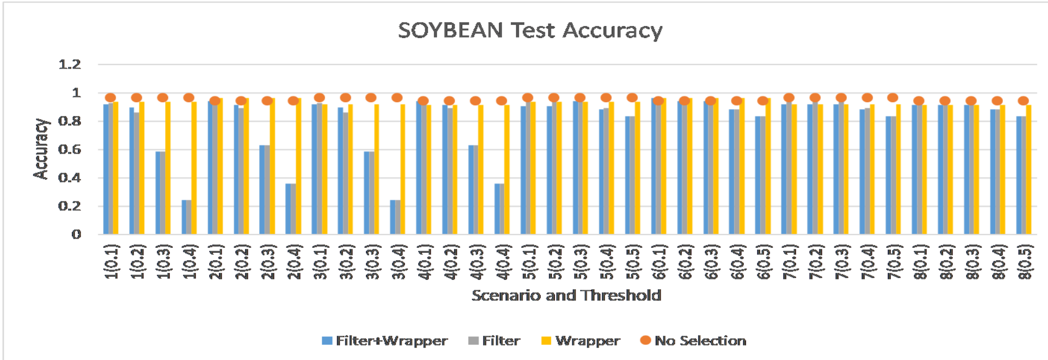


(c) Execution Time

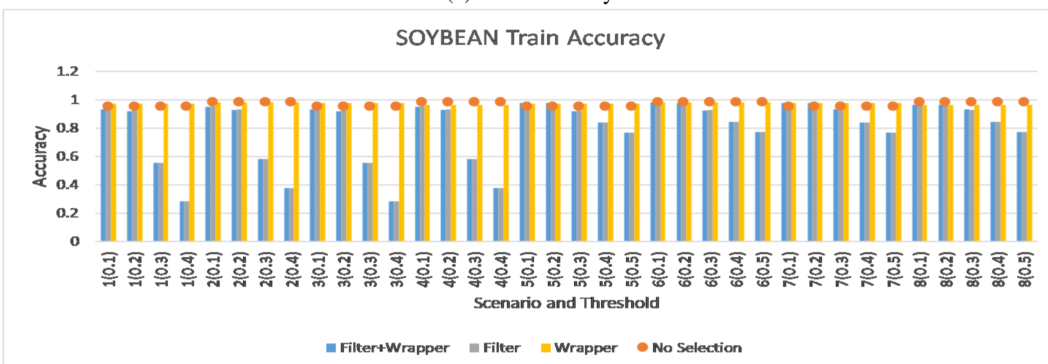


(d) Feature Subset Size

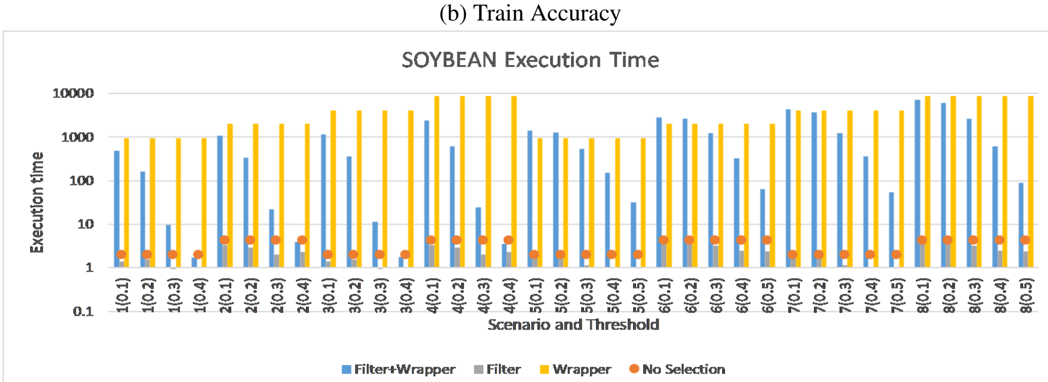
Figure 7: IONOSPHERE Experiment Data



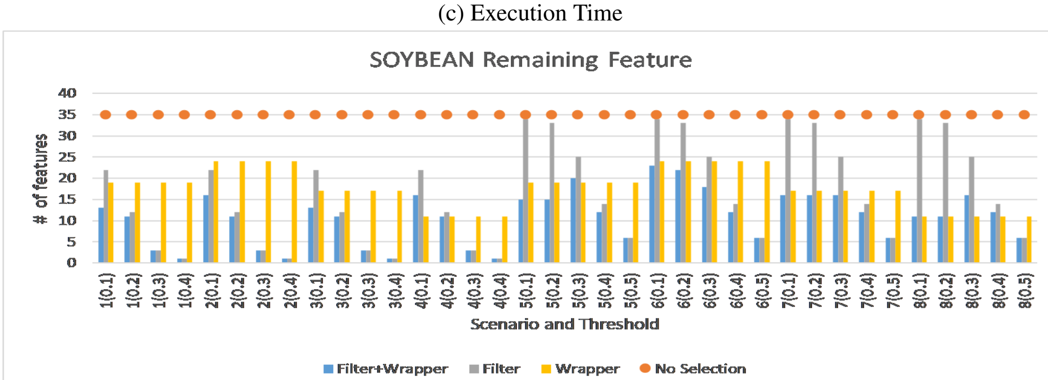
(a) Test Accuracy



(b) Train Accuracy

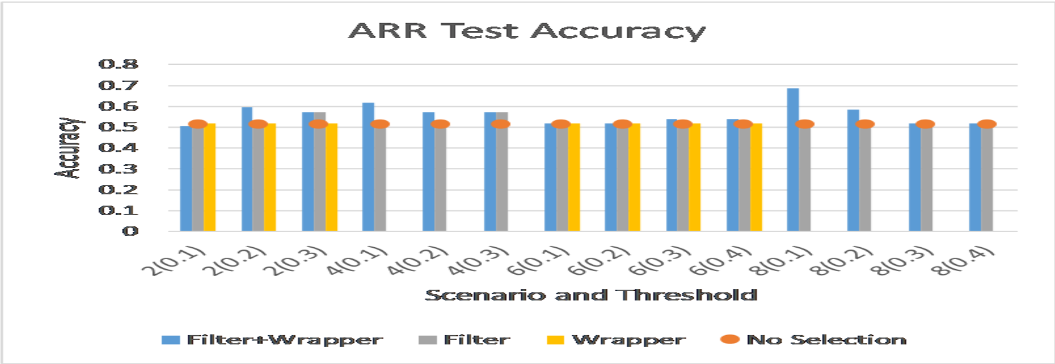


(c) Execution Time

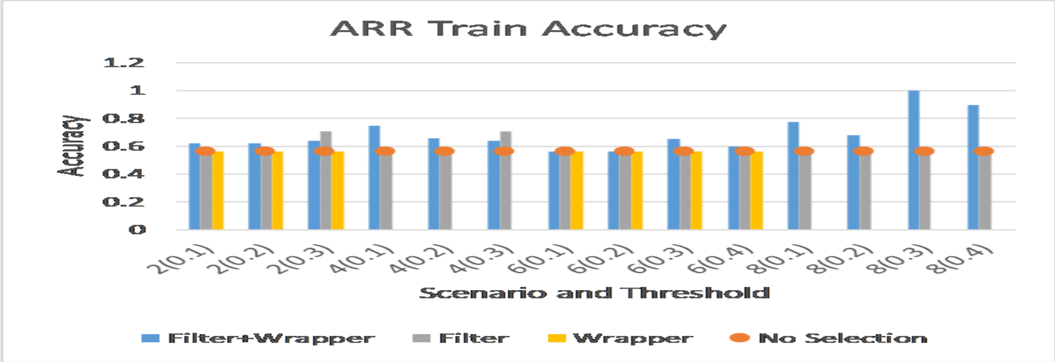


(d) Feature Subset Size

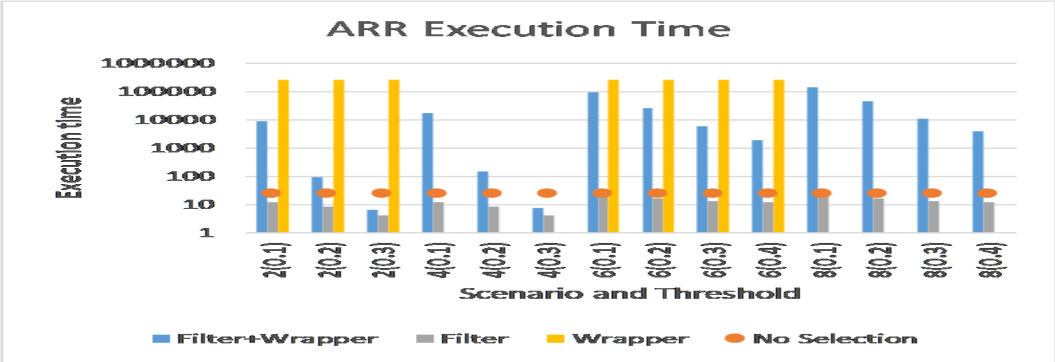
Figure 8: SOYBEAN Experiment Data



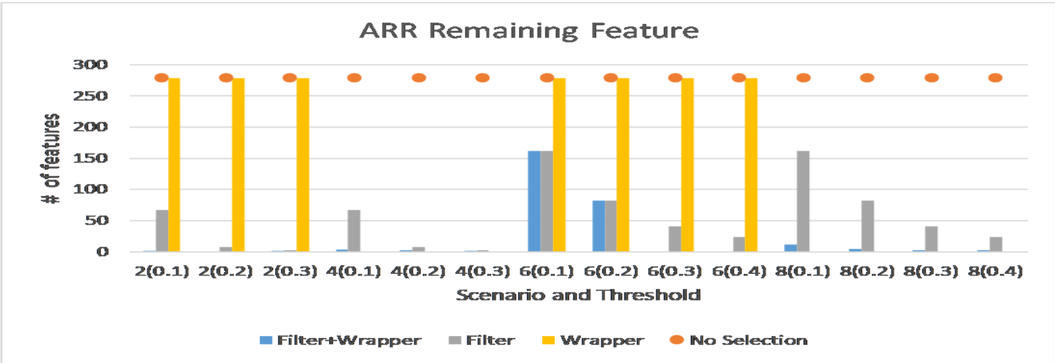
(a) Test Accuracy



(b) Train Accuracy

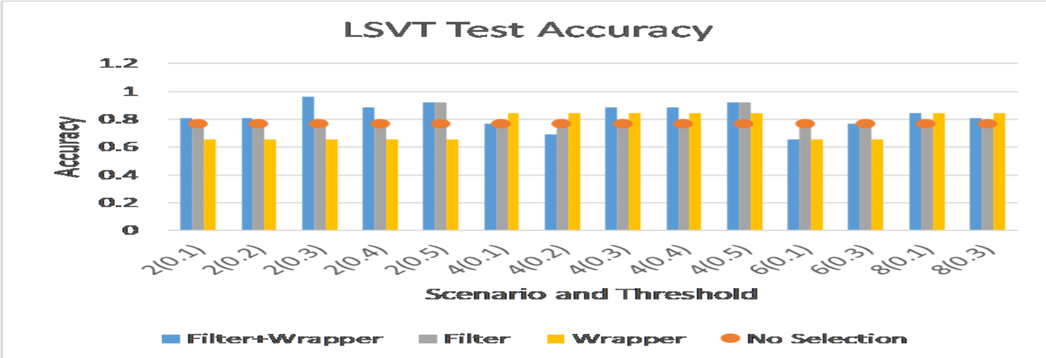


(c) Execution Time

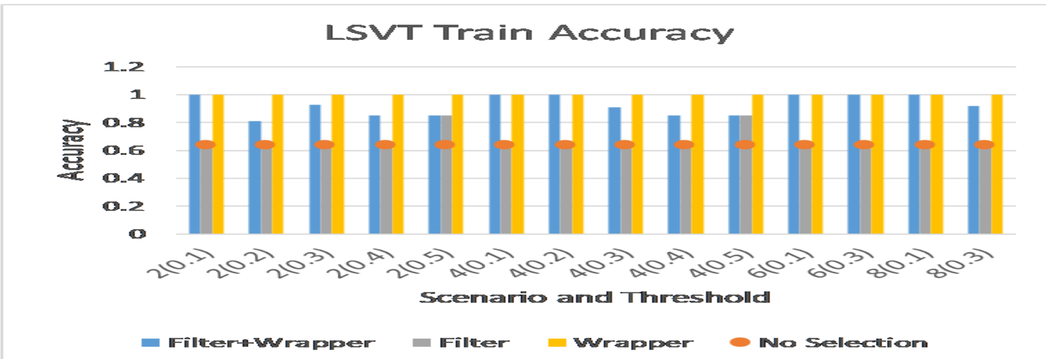


(d) Feature Subset Size

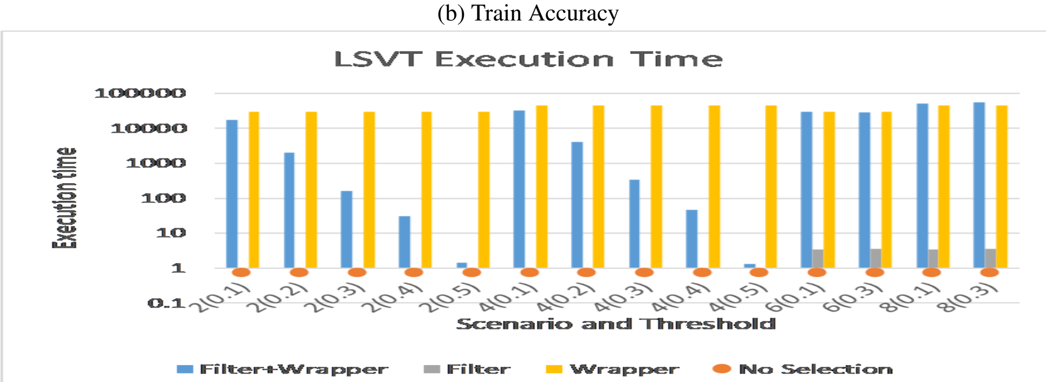
Figure 9: ARR Experiment Data



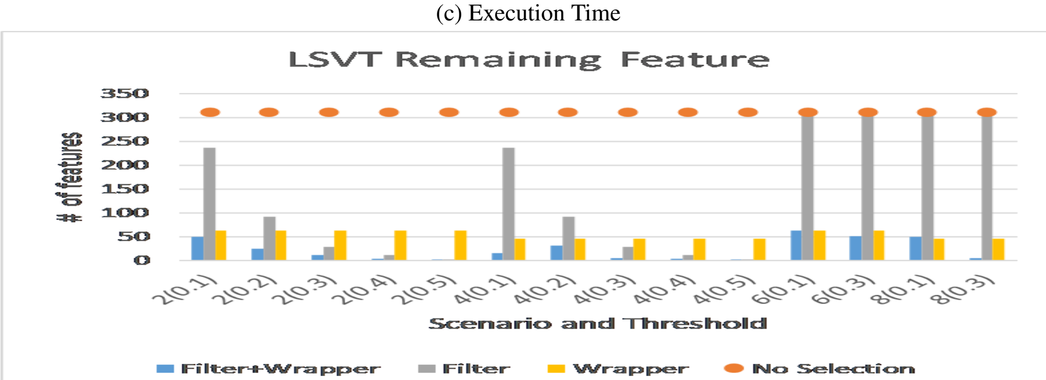
(a) Test Accuracy



(b) Train Accuracy

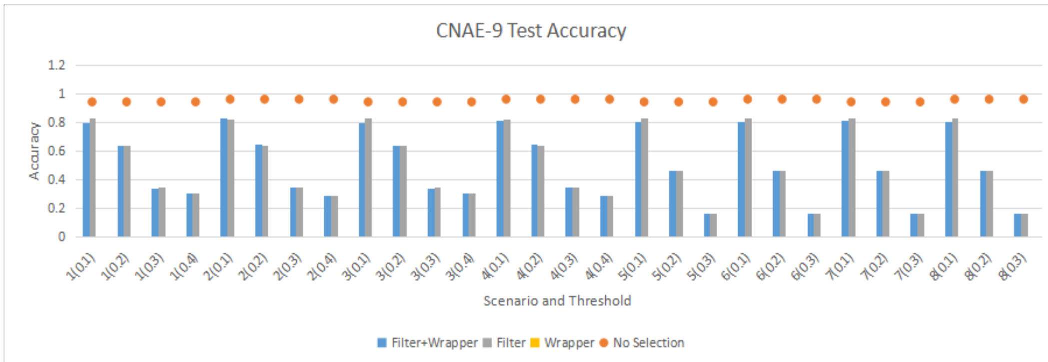


(c) Execution Time

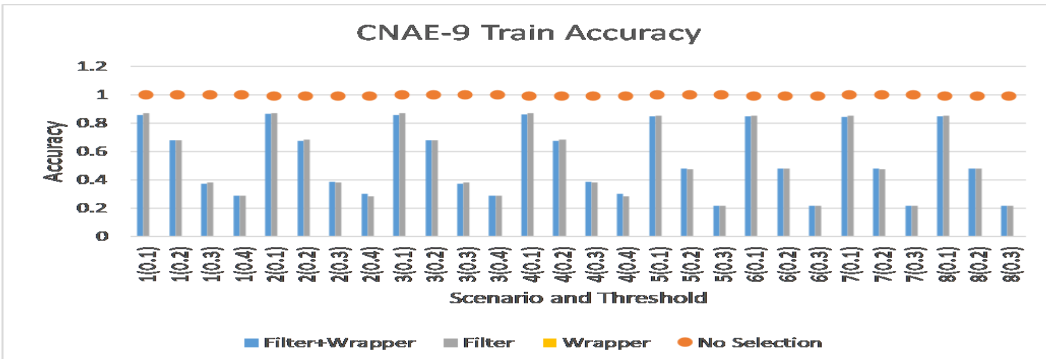


(d) Feature Subset Size

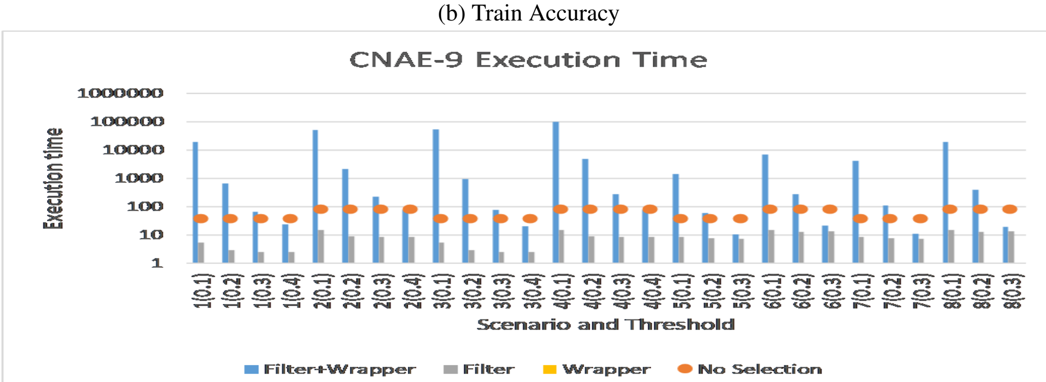
Figure 10: LSVT Experiment Data



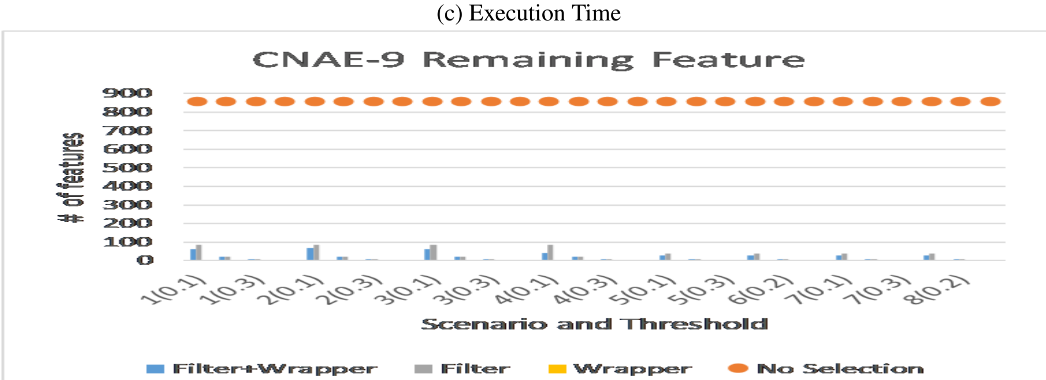
(a) Test Accuracy



(b) Train Accuracy



(c) Execution Time



(d) Feature Subset Size

Figure 11: CNAE-9 Experiment Data