ORIGINAL RESEARCH

Hybrid approaches to feature subset selection for data classification in high-dimensional feature space

Maysa Ibrahem Almulla Khalaf*[1,2], John Q Gan[1]

[1] School of Computer Science and Electronic Engineering, University of Essex, Essex, United Kingdom
[2] College of Science, Computer Science Department, University of Baghdad, Baghdad, Iraq

Received: July 18, 2019 **Accepted:** April 24, 2020 **Online Published:** September 22, 2020
DOI: 10.5430/air.v9n1p45 **URL:** https://doi.org/10.5430/air.v9n1p45

ABSTRACT

This paper proposes two hybrid feature subset selection approaches based on the combination (union or intersection) of both supervised and unsupervised filter approaches before using a wrapper, aiming to obtain low-dimensional features with high accuracy and interpretability and low time consumption. Experiments with the proposed hybrid approaches have been conducted on seven high-dimensional feature datasets. The classifiers adopted are support vector machine (SVM), linear discriminant analysis (LDA), and K-nearest neighbour (KNN). Experimental results have demonstrated the advantages and usefulness of the proposed methods in feature subset selection in high-dimensional space in terms of the number of selected features and time spent to achieve the best classification accuracy.

Key Words: Feature subset selection, Dimensionality reduction, Document categorization, Data mining

1. INTRODUCTION

In the past decades, the number of samples or observations and the number of features available for data mining have increased significantly in different applications, such as text categorization and document retrieval. Feature subset selection or dimensionality reduction is a very important processing step in document and text categorization and pattern recognition in general in high-dimensional feature space.[1–3]

Interpretability in data mining is an important issue in many fields such as social sciences and medicine.[4] Many traditional methods for dimensionality reduction can achieve high accuracy in data classification, but their results are usually difficult to interpret. For example, PCA is a commonly used method for feature extraction and dimensionality reduction, however, it combines original data into new features which are difficult to interpret.

Feature subset selection finds a subset of features with high predictive power and improved generalisation ability by reducing the chance of overfitting in subsequent data modelling and classification. In general, there are three methods for the performance evaluation of potential feature subsets: filter approach, wrapper approach, and embedded approach. Wrapper methods can be impractical when the number of features available for selection and the number of samples are too large, whilst the computational cost of filter methods is much less than wrapper methods for large feature datasets. However, wrapper methods are usually more accurate than filter methods.[5–11] Embedded approach searches locally for features that allow better local discrimination. It uses independent criteria to decide on the optimal subsets for given

*Correspondence: Maysa Ibrahem Almulla Khalaf; Email: miabdu@essex.ac.uk; Address: School of Computer Science and Electronic Engineering, University of Essex, Essex, United Kingdom.

cardinality, in which learning algorithms are usually used to select the final optimal feature subset. Embedded approach interacts with learning algorithms at a lower computational cost than wrapper approach.[6, 10, 13]

There are many different criteria applied to filter-based feature selection and dimensionality reduction, such as distance or similarity/dissimilarity criteria,[12, 13] information theory measures, and statistics measures. Distance or similarity/dissimilarity criteria have been applied for feature selection in many application areas, such as pattern recognition, information retrieval and detection of phishing emails and websites.[14] However, these measures are easily affected by noise or outlier data. Information theory measures have been widely used for feature selection, such as information gain (IG).[17–19] Recently, ensemble feature selection approach[20–23] has received much attention, which is commonly used for combining multiple models or methods to form a single effective method.

This paper proposes two hybrid methods for feature subset selection, consisting of two stages to select a relevant subset of features. The first stage selects feature subsets based on the union or intersection of features selected according to a variety of distance or similarity measures (unsupervised) mutual information measures (supervised). The second stage employs a wrapper approach on the selected features to further reduce the feature dimensionality and hopefully improve classification accuracy as well. Experiments were conducted with the performance of the proposed methods evaluated by comparison with the individual filter approaches and the full wrapper approach. This paper is organized as follows: Section 2 describes the basic principles of filter approaches based on both supervised and unsupervised criteria, the wrapper approach, and the proposed approach. Section 3 presents the experimental results and discussions. Conclusions are drawn in Section 4.

2. Filter approach, wrapper Approach and the proposed approach

2.1 Filter approach

A filter method selects a subset of features or ranks features based on some general characteristics of the features, independently without including any classification methods.[5–11] There are two main types of filter-based feature selection: unsupervised and supervised. Unsupervised methods select features according to distance or similarity/dissimilarity between features,[13, 14] whilst supervised methods select features according to their correlation or relevance with class labels.

1) Unsupervised Filter Approaches

• Euclidean Distance[24]

$$Euc\_dis(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

• Hamming Distance[25]

$$Ham\_dis(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}(x_i \neq y_i) \tag{2}$$

• City Block Distance[13, 26, 27]

$$City-block\_dis(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}|\, x_i - y_i \,| \tag{3}$$

• Hausdorff Distance[28]

$$H(\mathbf{x},\mathbf{y}) = \max(h(\mathbf{x},\mathbf{y}), h(\mathbf{y},\mathbf{x})) \tag{4}$$

where

$$h(\mathbf{x},\mathbf{y}) = \max_{x_i \in \mathbf{x}} \min_{y_i \in \mathbf{y}} |\, x_i - y_i \,| \tag{5}$$

2) Supervised Filter Approaches

When class information is available, various evaluation criteria for feature selection can be defined based on information theory. The two criteria adopted in this paper are described below.

• Information Gain (IG)

For a given feature set S that contains features from c classes, information gain (IG) from feature or attribute a is defined as follows:[29]

Given two vectors of features, $x = x_1, x_2, \ldots, x_n$ and $y = y_1, y_2, \ldots, y_n$, where n is the number of observations, various evaluation criteria for feature selection can be defined without using the corresponding class information, such as those defined below.

$$I(S,a) = I(S) - \sum_{v \in a} \frac{|\, S_{a,v} \,|}{|\, S \,|} I(S_{a,v}) \tag{6}$$

where v is a value of the attribute a, S_a,v is the subset of instances whose a has value v, $|S|$ is the number of instances in set S, and $I(S)$ denotes the entropy of feature set S, which is defined as

$$I(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{7}$$

where p_i is the percentage of features that belong to class i.

• Minimum Redundancy and Maximum Relevance (mRMR)

Minimum redundancy and maximum relevance is a criterion based on mutual information. Peng et al.[19] proposed this two-stage feature selection method to select a compact set of relevant features and they applied it to different datasets (handwritten digits, arrhythmia, NCI cancer cell lines and lymphoma).

Suppose m-1 features have already been selected from the available set of features S, forming a selected feature subset Sm-1. In order to select the next best feature, the mRMR method optimises the following condition:[5, 19]

$$\max_{x_j \in S-S_{m-1}} [I(x_j;\ c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j;\ x_i)] \tag{8}$$

where c represents class label and $I(x, y)$ is the mutual information function defined in terms of the joint probability of x and y and their marginal probabilities as follows:

$$I(x,y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \mathrm{d}x\mathrm{d}y \tag{9}$$

By maximizing the mRMR value, the method chooses the next feature that has maximum relevance to the class label and minimum redundancy to previously selected features.

2.2 Wrapper approach

Wrapper approach is a very common technique for feature subset selection, in which classifiers usually built up via machine learning are used for the evaluation of selected feature subsets, aiming to improve the classification performance.[5, 14, 29] However, it can be unrealistic when the number of features available for selection and the number of samples are too large, especially when the classifier training is computationally expensive. This approach may also suffer from overfitting.

The wrapper approach starts from a given subset G_0 which can be an empty set, a full subset, or any randomly selected subset. It then searches through the feature space using one of the search strategies suitable to this purpose. Subsequently, it evaluates each generated subset G_i by applying a learning model to the data labelled with G_i. If the performance of the learning model using G_i becomes better, G_i is considered to be the most recent best subset. For that reason, the wrapper approach then modifies G_i by adding or removing features to or from G_i (as dictated by the learning model) and the search iteration continues until a predefined stopping criterion is achieved.[29]

2.3 The proposed approach

Using a specific classifier, the wrapper approach compares cross-validation classification accuracies obtained with the potential feature subsets. Wrapper-based feature selection is vulnerable to overfitting due to its comprehensive search of the feature space and evaluation by a classifier constructed by machine learning. Hence, seeking reliable features using the wrapper method is sometimes impossible for datasets with a large feature space. In order to take the advantage of this highly accurate method and also to reduce its computational cost, hybrid approaches combining the advantages of the filter approach and the wrapper approach have been developed in recent years with different motivations and different search methods.[4, 29, 30] The hybrid strategy is adopted in the following two methods proposed in this paper. In the first stage, union and intersection among features selected by six filter methods are constructed to reduce unrelated and redundant features before the application of the costly wrapper method. The second stage further reduces the feature space considerably by using the wrapper method.

As one of the baseline methods for comparison, the wrapper approach can be applied to the original full feature dataset. Figure 1 illustrates how the full wrapper approach works. Three well-known classifiers with different structures and classification mechanisms are used in this study: support vector machine (SVM),[32, 33] linear discriminant analysis (LDA) and K-nearest neighbour (KNN).[34, 35]

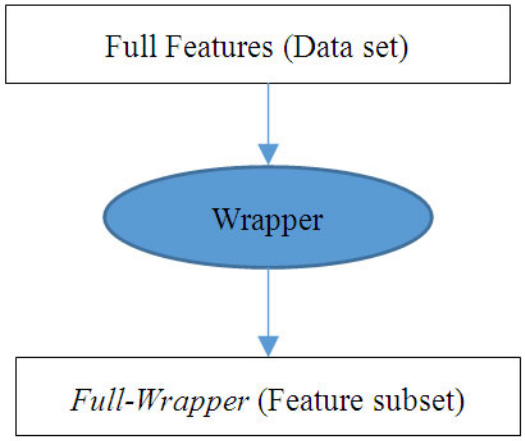

Figure 1. Wrapper approach

1) The Proposed Hybrid Approach 1

The first hybrid approach proposed in this paper employs four (unsupervised) distance or similarity measures, i.e., Euclidean, Hamming, City Block and Hausdorff, defined in (1), (2), (3) and (4) respectively, to compute the similarity between each pair of feature vectors in the training datasets,

generating four different similarity matrixes of the training data, which are used to select m most useful features. For each criterion, the optimal value of m is chosen by cross validation. The union of the four subsets of the selected features generates a combined feature subset called C1 at this first stage. In the second stage, the wrapper approach is applied to C1 to find an optimal feature subset H1. For comparison purposes, both C1 and H1 will be evaluated on test datasets. Figure 2 illustrates how this approach works.

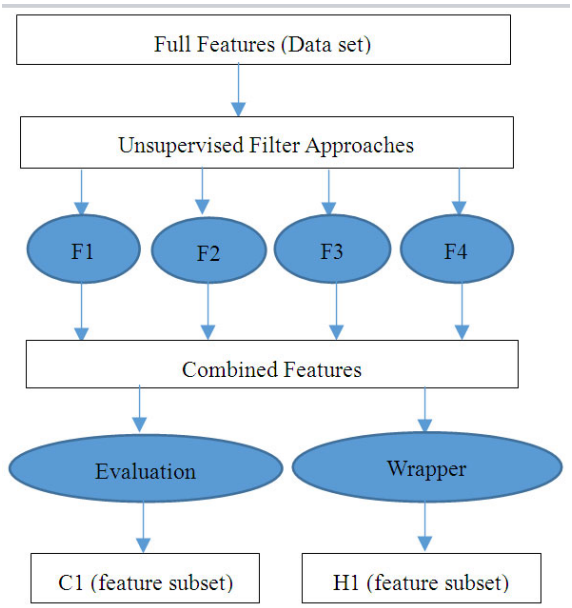

Figure 2. The proposed hybrid approach 1

2) The Proposed Hybrid Approach 2

The second hybrid approach proposed in this paper employs six (unsupervised and supervised) filter criteria. Four (unsupervised) distance or similarity measures, i.e., Euclidean, Hamming, City Block and Hausdorff, defined in (1), (2), (3) and (4) respectively, compute the similarity between each pair of feature vectors in the training datasets, generating four different similarity matrixes of the training data, which are used to select m most useful features. Two (supervised) mutual information based criteria IG and mRMR, defined in (6) and (8) respectively, select a compact set of relevant features from the training datasets, generating two different matrixes of the training datasets, which are used to select m most useful features. For each criterion, the optimal value of m is chosen by cross validation. The union of the two subset results, which are generated from the intersection of the four sets of features selected by unsupervised filters and the intersection of the two sets of features selected by supervised filters, forms a combined feature subset called C2 at this first stage. In the second stage, the wrapper approach is applied to C2 to find an optimal feature subset H2. For comparison purposes, both C2 and H2 will be evaluated on test datasets. Figure 3 illustrates how this approach works.

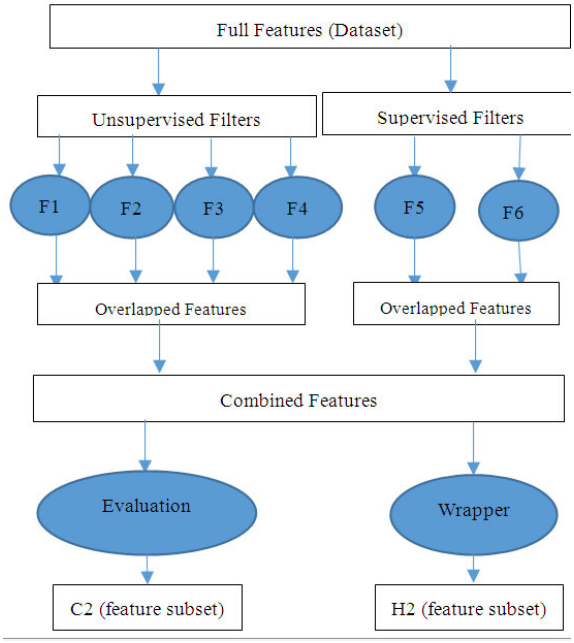

Figure 3. The proposed hybrid approach 2

3. Experimental Results and Discussions

3.1 Datasets

Seven datasets (four text and three numerical) were used in the experiments. First, an email dataset version 1(emails_v1) has 6,000 samples of two classes (3,000 ham/non-spam and 3,000 phishing) from different resources, such as Cornel University and Enron Company (`http://snap.stanford.edu/data/`). Second, the 20-newsgroup corpus dataset has approximately 20,000 newsgroup documents divided into 20 discussion groups (`https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups`). Because some newsgroups are very closely related to each other, four relatively distinguishable categories were used in the experiments in this study. Third, a document dataset from 10 categories of the Reuters-21578 dataset (`http://www.daviddlewis.com/resources/testcollection/reuters21578`) has 1,885 samples from 10 classes. Fourth, a musk dataset (`https://archive.ics.uci.edu/ml/datasets.html`) has 6,598 samples of two classes (musk and non-musk). Fifth, an email dataset version2 (emails_v2) (`https://www.kaggle.com/wcukierski/enron-email-datase`) has 1,000 samples from two classes (500 ham/non-spam and 500 phishing) from Cornel University and Kaggle Competition website. Sixth, a techni-

cal website features dataset (`http://khonji.org/phishing_studies`) has 4,230 samples from two classes (2,115 phishing and 2,115 non-phishing). Seventh, a self-drive intrusion detection dataset (`https://ll.mit.edu/ideval/data/2000data.html`) has 10,000 samples from two classes (5,000 malicious and 5,000 non-malicious). The number of features in these seven datasets are 1,014, 2,591, 412, 166, 465, 47, 80 respectively.

3.2 Experiment procedure

Each dataset, the experiment was repeated five times, with different data partition shuffled with different random seeds for each run in order to assess the consistency of the results. In each run the dataset was partitioned into a training set and a testing set. Part of the training set was used as validation data for choosing optimal parameter values for the various methods for comparison. The four text datasets (emails version 1&2, 20newsgroups, and Reuters) were pre-processed by tokenization and removing stop words such as 'the' and 'for' and numbers and symbols, which generated a bag of words (BOW) for each dataset as original features. After pre-processing, the total number of words for the emails_v1 dataset, emails_v2 dataset, 20 newsgroups dataset, and routers dataset was 1,014, 465, 2,591, and 412 respectively. After that, four term-weighting schemes were applied to the words in the BOW: term frequency (TF), term presence (TP), term frequency and inverse document frequency (TF-IDF),[36] and term presence and class-specific document frequency (TP-CSDF),[37] to generate numerical features. The other three datasets are numerical. The musk dataset has 166 features that describe properties of molecules, the technical dataset has 47 features that describe website technical features, and the malicious dataset has 80 features that describe malicious behaviour of self-drive vehicles. Finally, the proposed approaches were applied to select the optimal number of discriminate features in the sense that the highest cross-validation (5-folds) performance was achieved, and they were evaluated by employing three classifiers LDA, SVM, and KNN with the selected features on the test datasets.

3.3 Results

1) Classification accuracy and the corresponding number of required features: Figures 4–9 show the mean (over TF, TP, TF-IDF and TP-CSDF) of the classification accuracy and standard deviation on the testing datasets and the mean of the corresponding number of required features using LDA, SVM, and KNN respectively, by comparing among the six individual filter approaches, the combined filter approaches, the proposed hybrid approaches, and the full wrapper approach. The two proposed hybrid approaches achieved competitive accuracy with a significantly smaller number of required features compared to the full wrapper approach.

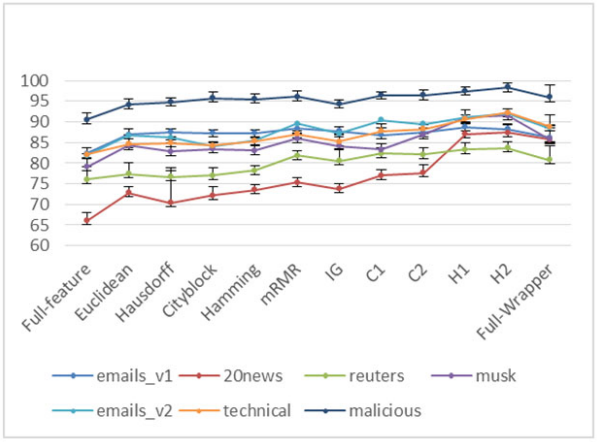

Figure 4. Accuracy and standard deviation of LDA with selected feature subsets on the seven datasets

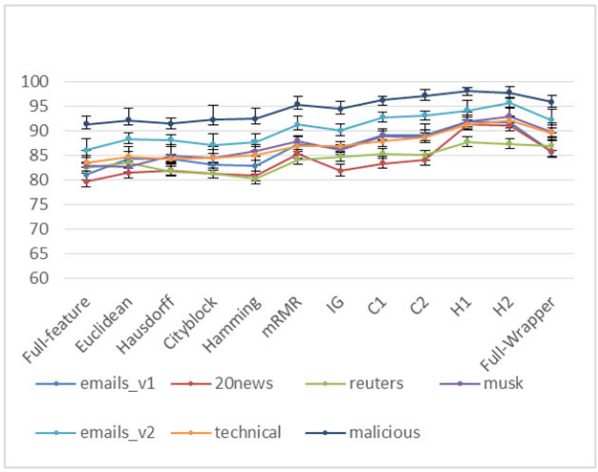

Figure 5. Accuracy and standard deviation of SVM with selected feature subsets on the seven datasets

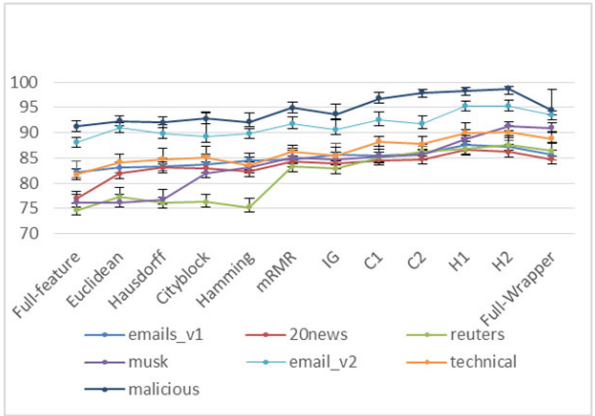

Figure 6. Accuracy and standard deviation of KNN with selected feature subsets on the seven datasets

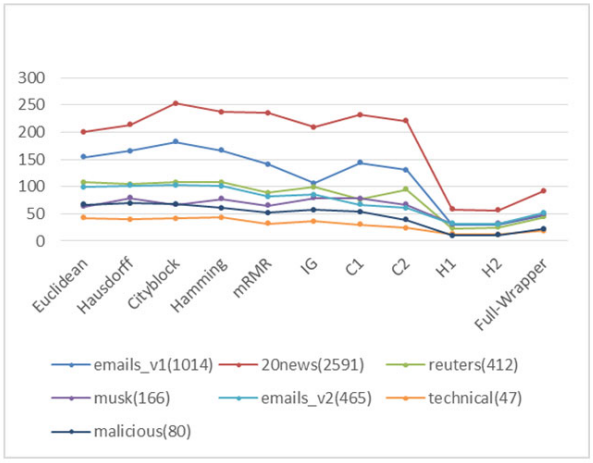

Figure 7. Number of selected features for LDA on the seven datasets

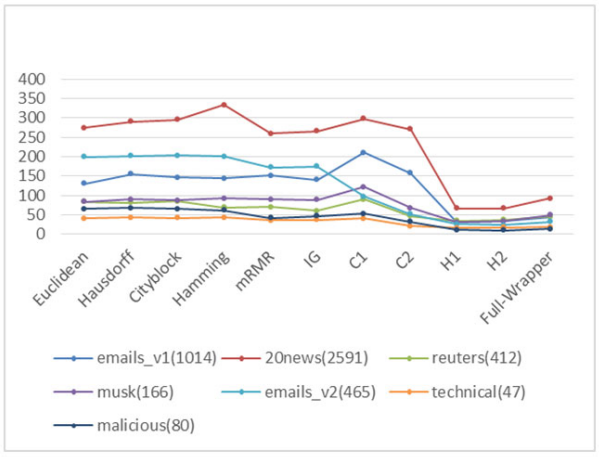

Figure 8. Number of selected features for SVM on the seven datasets

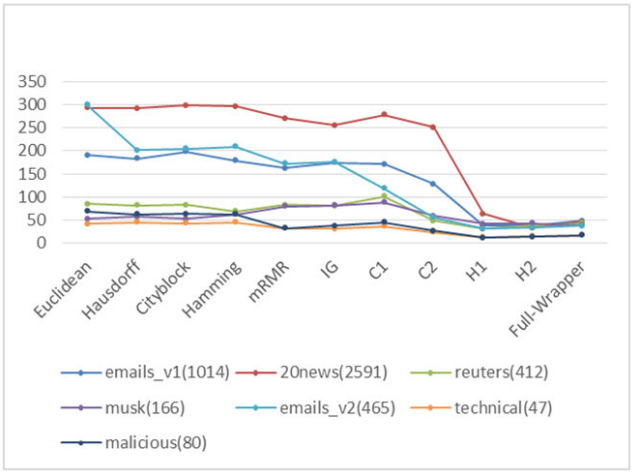

Figure 9. Number of selected features for KNN on the seven datasets

2) Computational time: Figures 10–12 demonstrate the mean of the time spent (in seconds) by various feature selection methods. As expected, the two proposed hybrid methods spent much less time than the full wrapper approach.

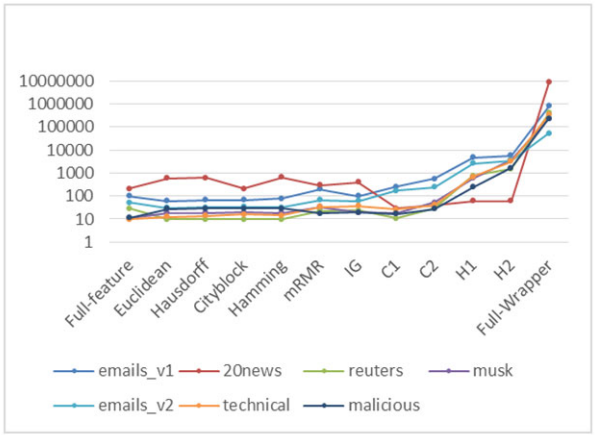

Figure 10. Time spent using LDA on the seven datasets

3) Interpretability of the selected features: Table 1 shows the top five terms in the emails v1 and v2 datasets for phishing detection, which correspond to the features selected by H1, H2, and Full-Wrapper respectively. As a matter of common sense, it seems that the top ten terms selected by the two proposed hybrid approaches are more relevant to phishing, such as 'bank' and 'click'. Similarly, better interpretability of the features selected by the proposed hybrid approaches was also observed on the news topics datasets. It seems that the terms (features) selected by the hybrid approaches are more interpretable than those selected by the full wrapper approach.

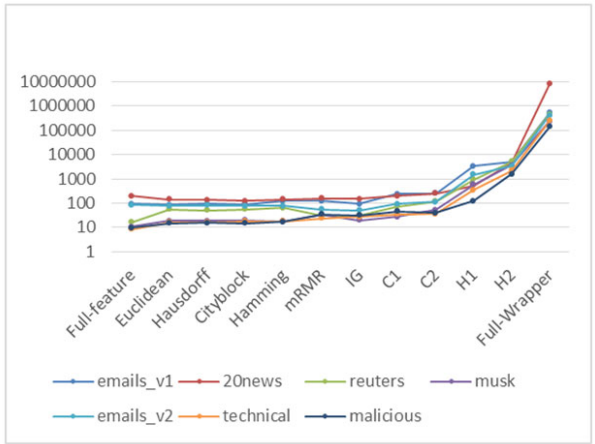

Figure 11. Time spent using SVM on the seven datasets

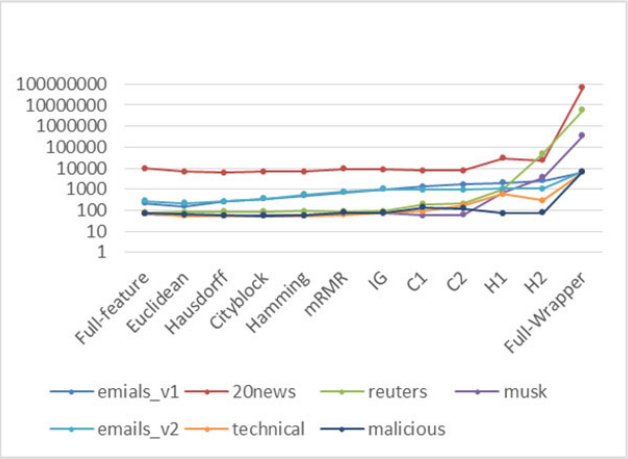

Figure 12. Time spent using KNN on the seven datasets

Table 1. Top Five Terms Selected from the Emails Dataset

H1	H2	Full-Wrapper
Attached	Online	Shop
Bank	Update	Original
Online	Attached	Effective
Click	Deal	Resources
www	Link	Company

Table 2. Statistical Test Results (t-test)

Method *Pair*	Accuracy	No. of required features	Time consumed
H1 vs C1	P = 0.0783	P = 2.3298e-06	P = 0.5421
H1 vs C2	P = 0.0423	P = 7.162e-05	P = 0.164
H2 vs C1	P = 0.0258	P = 3.7452e-06	P = 0.2746
H2 vs C2	P = 0.0364	P = 7.8182e-05	P = 0.3567
H1 vs Full	P = 0.345	P = 4.3362e-07	P = 4.2516e-07
H2 vs Full	P = 0.4765	P = 6.4249e-05	P = 3.1142e-07
H1 vs H2	P = 0.6385	P = 0.2423	P = 0.0050

Table 3. Statistical Test Results (rank-sum)

Method *Pair*	Accuracy	No. of required features	Time Consumed
H1 vs C1	P = 0.0034	P = 2.4321e-05	P = 0.6121
H1 vs C2	P = 0.0043	P = 2.2456e-05	P = 0.3151
H2 vs C1	P = 0.0307	P = 2.3287e-04	P = 0.5621
H2 vs C2	P = 0.0347	P = 3.2567e-04	P = 0.3421
H1 vs Full	P = 0.314	P = 3.3546e-06	P = 2.3467e-04
H2 vs Full	P = 0.4123	P = 4.3567e-04	P = 2.1361e-04
H1 vs H2	P = 0.3566	P = 0.2135	P = 0.0031

4) Statistical significance test: In order to assess whether the performance differences among the methods are statistically significant, we applied t-test, a parametric method, and Wilcoxon's rank-sum test, a non-parametric method, to determine whether two sets of performance data are significantly different from each other. The statistical tests were conducted to compare H1 against C1, H1 against C2, H2 against C1, H2 against C2, H1 against Full-Wrapper, H2 against Full-Wrapper, and H1 against H2, in terms of classification accuracy, time consumed, and the number of selected features. Tables 2 and 3 show the p-values for these pair comparisons, which demonstrate that H1 and H2 significantly outperformed C1 and C2 in terms of classification accuracy and the number of selected features, and H1 and H2 outperformed Full-Wrapper in terms of the time consumed and the number of selected features. In addition, H1 and H2 sometimes achieved higher accuracy than Full-Wrapper, but the difference is not statistically significant. Finally, H1 consumed significantly less time than H2.

4. Conclusions

This paper proposes two hybrid approaches to feature subset selection based on the combination of unsupervised and supervised filter approaches and the wrapper approach for data classification in high-dimensional feature space. They were tested on seven datasets from different resources and with different properties. Preliminary experimental results have demonstrated the advantages of the proposed methods over individual filter approaches and the full wrapper approach as well in terms of classification accuracy, the number of required features, consumed time, and interpretability. Furthermore, the first hybrid approach H1 is better than the second approach H2 in terms of time consumed, but H2 outperforms H1 in terms of classification accuracy.

We observed that SVM is vulnerable to over-fitting with the wrapper approach working on full features. This can be illustrated by the fact that the non-linear classifier with the wrapper method did not achieve satisfactory testing accuracy, particularly with complicated data space.

The stability of the selected features is desirable in practical feature selection applications,[25] which has not been investigated yet in this paper. It is also desirable to compare with more other hybrid feature selection methods, such as those proposed in[8] and,[26] to further evaluate the proposed methods. These would be considered in our future work in this line of research.

Conflicts of Interest Disclosure

The authors declare that they have no competing interests.

REFERENCES

[1] Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of ICML, 2003, pp. 856-863.

[2] Muazzam AS. An empirical evaluation of text classification and feature selection methods. Artificial Intelligence Research. 2016; 5: 70-81. https://doi.org/10.5430/air.v5n2p70

[3] Ibrahim MA, Mukundan R. Cascaded techniques for improving emphysema classification in computed tomography images. Artificial Intelligence Research. 2015; 4: 112-114. https://doi.org/10.5430/air.v4n2p112

[4] Katuwal GJ, Chen R. Machine learning model interpretability for precision medicine. arXiv preprint. 2016: 1610-1614.

[5] Gan JQ, Hasan BAS, Tsui CSL. A filter-dominating hybrid sequential forward floating search method for feature subset selection in high-dimensional space. International Journal of Machine Learning. 2014; 5: 413-423. https://doi.org/10.1007/s13042-012-0139-z

[6] Dash M, Liu H. Feature selection for classification. Intelligent Data Analysis. 1997; 1: 131-156. https://doi.org/10.3233/IDA-1997-1302

[7] Quoc HB. A combined approach for filter feature selection in document classification. In: IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI), 2015, pp. 317-324.

[8] Zhen Z, Wang H, Xing Y, et al. Text feature selection approach by means of class difference. In: International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016, pp. 1582-1586.

[9] Dhote Y, Agrawal S, Deen AJ. A survey on feature selection techniques for Internet traffic classification. In: International Conference on Computational Intelligence and Communication Networks (CICN), 2015, pp. 1375-1380.

[10] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research. 2003; 3: 1157-1182.

[11] Tang J, Alelyani S, Liu H. Feature selection for classification: A review. Data Classification: Algorithms and Applications. 2014: 3-27.

[12] Lu, Y, Liu W, He X. A text feature selection method based on category-distribution divergence. Artificial Intelligence Research. 2015; 4(2): 143-148. https://doi.org/10.5430/air.v4n2p143

[13] Cha SH. Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences. 2007; 1: 300-307.

[14] Saeys Y, Inza I, Larranga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23: 2507–2517. PMid:17720704. https://doi.org/10.1093/bioinformatics/btm344

[15] Vega R, Sajed T, Mathewson KW, et al. Assessment of feature selection and classification methods for recognizing motor imagery tasks from electroencephalographic signals. Artificial Intelligence Research. 2017; 6(1): 37-51. https://doi.org/10.5430/air.v6n1p37

[16] Tian XP, Geneg GG, Li HT. A framework for multi-features based web harmful information identification. In: Proceedings of International Conference on Computer Application and System Modelling, 2010, pp. V11-614–V11-618.

[17] Chandrasekaran M, Narayanan K, Upadhyaya S. Phishing email detection based on structural properties. In: Proceedings of the NYS Cyber Security Conference, 2006, pp. 1-7.

[18] Gomez JC, Boiy E, Moens MF. Highly discriminative statistical features for email classification. Knowledge and Information Systems. 2012; 31: 23–53. https://doi.org/10.1007/s10115-011-0403-7

[19] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and minredundancy. IEEE Transactions on Pattern Analysis and MachineIntelligence. 2005; 27: 1226-1238. PMid:16119262. https://doi.org/10.1109/TPAMI.2005.159

[20] Fahad A, Tari Z, Khalil I, et al. An optimal and stable feature selection approach for traffic classification based on multi-criterion fusion. Future Generation Computer System. 2014; 36: 156-169. https://doi.org/10.1016/j.future.2013.09.015

[21] Wang H, Khoshgoftaar MT, Napolitano A. Software measurement data reduction using ensemble techniques. Neurocomputing. 2012; 92: 124-132. https://doi.org/10.1016/j.neucom.2011.08.040

[22] He Z, Yu W. Stable feature selection for biomarker discovery. Computational Biology and Chemistry. 2010; 34: 215-225. PMid:20702140. https://doi.org/10.1016/j.compbiolchem.2010.07.002

[23] Aldehim GN. Heuristic Ensembles of Filters for Accurate and Reliable Feature Selection. PhD Thesis, University of East Anglia. 2016.

[24] Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In: Proceedings of ICML, 1997, pp. 412-420.

[25] Kalbhor M, Shrivastava S, Ujjainiya B. An artificial immune system with local feature selection classifier for spam filtering. In: Proceedings of the 4th International Conference on Computing, Communications and Networking Technologies, 2013, pp. 1-7.

[26] Piramuthu S. Evaluating feature selection methods for learning in data mining applications. European Journal of Operational Research. 2004; 156: 483-494. https://doi.org/10.1016/S0377-2217(02)00911-6

[27] Kano E, Tsuda K. Use of a text mining method for classification citizen report data and analyzing the occurance trend of local problems. Artificial Intelligence Research. 2019; 8(2): 61-70. https://doi.org/10.5430/air.v8n2p1

[28] Sikora R, Piramuthu S. Efficient genetic algorithm based data mining using feature selection with Hausdorff distance. Information Technology and Management. 2005; 6: 315-331. https://doi.org/10.1007/s10799-005-3898-3

[29] Karegowda AG, Jayaram M, Manjunath A. Feature subset selection problem using wrapper approach in supervised learning. International Journal of Computer Applications. 2010; 1: 13-17. https://doi.org/10.5120/169-295

[30] Das S. Filters, wrappers and a boosting-based hybrid for feature selection, In: Proceedings of ICML, 2001, pp. 74-81.

[31] Zhu Z, Ong YS, Dash M. Wrapper-filter feature selection algorithm using a memetic framework. IEEE Transactions on Systems, Man and Cybermetics, Part B. 2007; 37: 70-76. PMid:17278560. https://doi.org/10.1109/TSMCB.2006.883267

[32] Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and Computing. 2004; 14: 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

[33] Vapnik V. The Nature of Statistical Learning Theory, Springer Science & Business Media. 2000.

[34] Altman NS. An introduction to kernel and nearest-neighbour nonparametric regression. The American Statistician. 1992; 46: 175–185. https://doi.org/10.1080/00031305.1992.10475879

[35] Anne C, Mishra A, Hoque MT, et al. Multiclass patent document classification. Artificial Intelligence Research. 2018; 7(1): 1-14. https://doi.org/10.5430/air.v7n1p1

[36] Plansangket S, Gan JQ. A query suggestion method combining TF-IDF and Jaccard Coefficient for interactive web search. Artificial Intelligence Research. 2015; 4: 119-125. https://doi.org/10.5430/air.v4n2p119

[37] Plansangket S, Gan JQ. Term presence and class specific document frequency for document representation and classification. In: Proceedings of the 7th Computer Science and Electronic Engineering Conference, UK, 2015, pp. 5-8. https://doi.org/10.1109/CEEC.2015.7332690