Article:
Mustapa, N.A., Senawi, A. and Wei, H.-L. (2023) Supervised feature selection based on the law of total variance. MEKATRONIKA, 5 (2). pp. 100-110. ISSN: 2637-0883

https://doi.org/10.15282/mekatronika.v5i2.9998

MEKATRONIKA
e-ISSN: 2637-0883
VOL. 5, ISSUE 2, 100 – 110
DOI: https://doi.org/10.15282/mekatronika.v5i2.9998

ORIGINAL ARTICLE

Supervised Feature Selection based on the Law of Total Variance

Nur Atiqah Mustapa[1], Azlyna Senawi[1,*] and Hua-Liang Wei[2]

[1]Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuhraya Persiaran Tun Khalil Yaakob, 26300 Gambang, Kuantan, Pahang, Malaysia.
[2]Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD, United Kingdom.

ABSTRACT – Feature selection is an essential pre-processing phase in machine learning that decreases data dimensionality by removing superfluous and irrelevant features. This paper presents a supervised method for selecting significant features that makes use the law of total variance (LTV). The LTV is specifically employed to quantify feature relevancy by evaluating the correlation of each feature with the class label. The proposed feature selection method was tested on eleven public datasets and six distinct classifiers to assess its performance and reliability. Findings show that the feature subset obtained by the LTV may achieve comparable classification accuracy as the whole feature set, even when only 50% or less than 50% of the original features are retained. The LTV was also proven versatile as it can achieve adequate classification accuracy with all six classifiers with different learning schemes. In addition, a comparison with a similar type of feature selection method (AmRMR) shows that the LTV performed a superior accuracy in classification.

ARTICLE HISTORY
Received: 17th Nov 2023
Revised: 24th Nov 2023
Accepted: 18th Dec 2023
Published: 28th Dec 2023

KEYWORDS
Correlation measure
Dimensionality reduction
Feature selection
Law of total variance
Classification

INTRODUCTION

In the past, databases were limited since data was only gathered in response to human requests to store and retrieve information. As the capacity to generate and store data has evolved fast in recent years, data capacity has grown by about 50% monthly regarding data types, volumes, locations, and currency [1]. Big data has led to an overabundance of data storage, processing, and reporting breakthroughs, enabling in-depth examination of all available data. As a result, big data has become a platform that facilitates extraordinary growth in data mining. Robust model development and flexible pattern finding are necessary to extract relevant information from the enormous data set. Real-world data often consists of vast quantities of unprocessed data with flaws such as noise, duplication, and superfluous representation. These forms of data are considered to be of low quality, potentially leading to a decline in data mining performance. As a result, a pre-processing step is necessary since it enables the data to be given in the ideal quantity, structure, and format for the best data mining outcomes.

Data reduction is a pre-processing step that produces new data by reducing the amount of original data without altering the original dataset's core structure [2]. There are three data reduction techniques: cardinality reduction, sample numerosity reduction, and dimensionality reduction. Cardinality reduction is implemented when data reduction applies data transformations such as the binning process to create a more compact representation of the original data overlap. Sample numerosity reduction replaces the original data with a model approximation and is equivalent to the original data but consumes less space than the actual data display. Regression and log-linear models are examples of sample numerosity reduction methods [3].

Dimensionality reduction is a widely employed technique aimed at reducing the number of features of a dataset. Limiting the dimensions of space and examining informative features is essential in many situations. There are two main approaches used for reducing dimensionality in data analysis: feature extraction and feature selection [2]. Feature extraction is a process that integrates and alters the original features to generate a new set of features. This new collection of features has smaller dimensions but still retains most of the relevant information in the dataset [4]. In contrast to feature extraction, feature selection aims to keep the originality of features by sustaining the essential information in the dataset for subsequent analysis while selecting only some of them to represent the full feature set.

Feature selection in supervised learning takes advantage of the presence of class labels. If a feature doesn't provide enough details for the intended class label, it may be eliminated. Recent examples of this method are supervised feature selection by constituting a basis for the original space of features and matrix factorization (SFS-BMF) [5], supervised filter feature selection method based on the spectral analysis and redundant analysis (RnR-SSFSM) [6] and self-learning multi-output regression (SPLR-FS) [7]. On the other hand, unsupervised feature selection is an approach where the class labels are not pre-established. In this approach, clustering techniques are often used to determine the composition of groups within a given dataset. Examples of unsupervised feature selection are the feature dependency-based unsupervised feature selection (DUFS) [8] and unsupervised feature selection by self-paced regularization (UFS-SP) [9].

*CORRESPONDING AUTHOR | Azlyna Senawi | ✉ azlyna@umpsa.edu.my

Semi-supervised feature selection involves the identification of significant features by identifying a subset of features that yields the most informative pattern through the combination of both labelled and unlabelled data. Examples of feature selection methods based on semi-supervised learning are sparse rescaled linear square regression (SRLSR) [10] and multi-view adaptive semi-supervised feature selection (MASFS) [11].

One of the commonly used criteria for feature selection is the correlation-based measure. Among the methods that employed this criterion are the integration of Relief-F with correlation-based feature selection (ReCFS) [12], hybrid correlation feature selection with genetic algorithm (CFGA) [13], adaptive correlation features selection and deep belief neural networks [14] and feature selection ordered by correlation (FSOC) [15]. The experimental findings have shown that the correlation-based criterion can serve as a distinct measure in selecting features, allowing for efficient execution of data mining tasks.

This study suggests a method for selecting features based on their relevance. In this regard, a feature relevance is evaluated by utilising the law of total variance (LTV). The LTV is based on the fundamental idea that the total variance can be calculated by combining the expected value of the conditional variance and the variance of the conditional means [16,17].

The subsequent sections of this paper are broken up as follows: the principle of the LTV is provided in Section 2, while Section 3 discusses the methodology of LTV. Section 4 provids the experimental procedures of this study. The findings from the experiment are analysed and explained in Section 5. Finally, Section 6 gives the conclusion of the paper by summarising the research content.

THE LAW OF TOTAL VARIANCE

This section discusses the derivation of the LTV equation, which is based on several essential principles. The discussion includes some mathematical justifications.

Variance

Let a quantitative random variable is denoted by G and a nominal random variable is denoted by H with potential outcomes of h_i. The definition of the marginal (overall) variance of G is:

$$Var(G) = E([G - E(G)]^2) = E(G^2) - [E(G)]^2. \tag{1}$$

Equation (1) demonstrates that variance is the square of the average deviation of G from its mean and is always a positive value. A low variance implies that G will likely have values closely clustered around the mean. A high variance implies that G exhibits substantial deviations from its mean.

According to Equation (1), the conditional variance for G given $H = h_i$ is

$$Var(G|H) = E(G^2|H) - [E(G|H)]^2. \tag{2}$$

Hence, the expected value of $Var(G|H)$ can be expressed using the linearity of expectation as follows:

$$E[Var(G|H)] = E[E(G^2|H)] - E([E(G|H)]^2). \tag{3}$$

The Law of Iteration Expectation (LIE)

The expected value or means of G, commonly referred to as the marginal expectation, is given as:

$$E(G) = \sum_{h_i} P(H = h_i)E(G|H = h_i) \tag{4}$$

where $E(G|H = h_i)$ is an expectation conditional on the h value. Modifying the value of h, will also result results in a change in the expression $E(G|H = h_i)$. Henceforth, the following is the equation for the expected value of conditional expectation, or often referred as LIE:

$$E(G) = E[E(G|H)]. \tag{5}$$

According to LIE, the total average $E(G)$ is calculated by taking the average of case-by-case averages $E[E(G|H)]$, where $E(G|H)$ is the individual averages of G given H for all possible h values that known as case-by-case averages.

Then, the variance of conditional expectation $\mathrm{Var}[E(G|H)]$ can be explicitly conveyed by using Equation (1) as shown in Equation (6):

$$Var[E(G|H)] = E([E(G|H)]^2) - (E[E(G|H)])^2. \tag{6}$$

The substitution of Equation (5) in Equation (6) denotes:

$$Var[E(G|H)] = E([E(G|H)]^2) - [E(G)]^2. \quad (7)$$

Interpretation of the LTV

Applying the first term of the right-hand side of Equstion (3) with Equation (5) gives:

$$E[Var(G|H)] = E(G^2) - E([E(G|H)]^2). \quad (8)$$

Combining Equation (8) with Equation (7) yields:

$$E[Var(G|H)] + Var[E(G|H)] = E(G^2) - [E(G)]^2. \quad (9)$$

Finally, applying Equation (1) to Equation (9) gives the equation that is recognised as the LTV:

$$Var(G) = E[Var(G|H)] + Var[E(G|H)]. \quad (10)$$

All terms in the Equation (10) must be in positive value since variance values are inherently positive [18]. The two terms that make up the LTV equation are: $E[Var(G|H)]$ and $Var[E(G|H)]$. The average variance of G over all potential outcomes is denoted by $E[Var(G|H)]$, where $Var(G|H = h_i)$ represents the variation of G in the potential outcome of H. This term that also known as the average within-sample variance refers to the variability of the average within outcomes [17]. On the other hand, $Var[E(G|H)]$ is the variance between the average of G over all potential outcomes, where $E(G|H = h_i)$ is the average of G over all the potential outcomes of H. This term measures the amount of variation between outcomes by quantifying the between-sample variance [17].

THE PROPOSED FEATURE SELECTION METHOD

1. **Input**: $D = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N\}$ // A complete dataset with N features
2. **Output**: S_{sig} // Subset of features
3. Normalize data if necessary
4. for $j = 1$ to N do begin
5. EVar = calculate $E[\text{Var}(\mathbf{f}_j|\mathbf{c})]$;
6. VarX = calculate $(\mathbf{f}_j)$;
7. Find $\text{LTV}_j = r_{\text{qn}}{}^2(\mathbf{f}_j, \text{c})$;
8. end for
9. S_{rel} = sorted and rank $\mathbf{f}_j$

Figure 1. Pseudocode of the proposed method.

The feature selection algorithm presented in Figure 1 consists of two major components. The initial step (line 5) is to assess the relevance of a feature by considering the class labels. Equation (8) is used to calculate the average variance of features in each class label, $EVar$, while Equation (1) is employed to determine the variance of features, $VarX$. The correlation between the feature and the class label, LTV_j, is subsequently assessed (line 7) using the criterion given in Equation (12). In the final step (line 9) of the algorithm, features are sorted, starting with the most relevant and ending with the least relevant features.

Monitoring Criterion Based on Feature Relevancy

This study employed the correlation-based measure to assess the relevance of a feature associated with the targeted class label $\mathbf{c}$. By utilizing the LTV given in Equation (10), the correlation between a feature $\mathbf{f}_j$ and the class label $\mathbf{c}$ is measured based on the following monitoring crierion:

$$r_{qn}(\mathbf{f}_j, \mathbf{c}) = \left(1 - \frac{E[Var(\mathbf{f}_j|\mathbf{c})]}{Var(\mathbf{f}_j)}\right)^{\frac{1}{2}}. \quad (11)$$

where $0 \leq r_{qn}(\mathbf{f}_j, \mathbf{c}) \leq 1$. There is likely a strong correlation between a feature $\mathbf{f}_j$ and the class label $\mathbf{c}$ when the value of $r_{qn}(\mathbf{f}_j, \mathbf{c})$ is close to '1'. The correlation between them is deemed as weak if the value $r_{qn}(\mathbf{f}_j, \mathbf{c})$ is close to '0'. The feature $\mathbf{f}_j$ and class label $\mathbf{c}$ exhibit perfect correlation if $r_{qn}(\mathbf{f}_j, \mathbf{c}) = 1$ while there is no correlation at all between them when $r_{qn}(\mathbf{f}_j, \mathbf{c}) = 0$ [19].

Hence, the relevance criterion can be denoted as below:

$$LTV_j = r_{qn}^2(\mathbf{f}_j, \mathbf{c}) \text{ such that } \mathbf{f}_j \in D. \quad (12)$$

The range of ${r_{qn}}^2$ is identical to r_{qn} since ${r_{qn}}^2$ employs the squared value of the r_{qn} from Equation (11) [16,18].

EXPERIMENTAL SETUP AND PROCEDURE

An experiment was done to evaluate and examine the effectiveness of the LTV method. A well-known data structure was employed first to test the efficacy of the LTV feature selection method in order to test how well the method works in choosing the most relevant features. The proposed method was then tested on eleven real public datasets and the results were compared with those obtained based on a criterion that considers both feature relevency and feature redundancy, called the advanced minimum redundancy maximum relevance (AmRMR) method [20].

Testing with Well-known Data Structure

Initially, a well-known data structure, the Iris dataset, was utilised to determine if the LTV method could choose and rank the features accurately according to their significance with the class labels. Numerous feature selection methods were tested and validated using this datasat. The dataset consists of 150 observations and includes four features: sepal length ($\mathbf{f}_1$), sepal width ($\mathbf{f}_2$), petal length ($\mathbf{f}_3$) and petal width ($\mathbf{f}_4$). There are no missing values in the dataset, and it is divided into three classes. There are 50 instances in each class, and the class labels are based on the Iris plant classes: Setosa, Versicolor, or Virginica. The Setosa class exhibits linear separability from the other two classes. However, the other two classes are not distinguishable from each other in a linear manner. It is broadly recognised that $\mathbf{f}_3$ and $\mathbf{f}_4$ are more vital features than the $\mathbf{f}_1$ and $\mathbf{f}_2$. Further discussion will be given in Section 5.1.

Testing with Unknown Data Structure

Eleven real public datasets [21] with unknown data structures were used to further assess the effectiveness of the proposed feature selection method. These datasets were selected based on three distinct categories of dimensional size: low, medium and high, which are $(N \leq 10)$, $(10 < N \leq 100)$ and $(N > 100)$, respectively. Table 1 summarises the main attributes of the datasets. Features with only one value will be discarded first because they will not show any significant correlation with the class label.

Table 1. Main Attributes of the tested datasets

Dataset	Number of features	Number of instances	Number of classes
Vertebral Column	6	310	3
Ecoli	9	336	8
Rice	10	18185	2
Statlog (Image)	19	2310	7
Mfeat Zernike	47	2000	10
100 Plant Species	64	1600	100
Mfeat Karhunen	64	2000	10
Mfeat Fourier	76	2000	10
Mfeat Factor	216	2000	10
Mfeat Pixel	240	2000	10
Isolet	617	7797	26

Comparison with the AmRMR Method

A comparison to the AmRMR method [20] was also conducted on each dataset presented in Table 1. The AmRMR method was chosen for comparison as it also applied a correlation-based measure to identify the best subset of features. The AmRMR method selects the best feature subsets based on the redundancy and relevance of the features. In contrast, the LTV method focuses solely on the feature relevance when selecting the best subset of features. The AmRMR method is relatively more complex compared to the LTV method as it uses a machine learning algorithm to determine feature relavance.

Validation Classifiers and Procedures

The datasets listed in Table 1 were employed to examine how well the LTV feature selection performed in terms of classification accuracy. The study also employed six classification algorithms, namely Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Classification and Regression Trees (CART), Bagging, and AdaBoost. According to the IEEE International Conference on Data Mining (ICDM), these six classifiers are widely recognised as among the most prominent data mining algorithms for various tasks [22]. The holdout cross-validation was applied for each classifier, where 80% and 20% of every dataset were allocated as the training set and the testing set, respectively. The KNN classifier utilises a fixed value of 5 for the number of nearest neighbours (k) for all datasets [23–25].

RESULTS AND DISCUSSION

The efficacy of the LTV for ranking significant features is first assessed based on the Iris dataset, and the results are compared with other established feature selection methods. This section provides a detailed description of evaluating the LTV's effectiveness compared to the AmRMR method based on the number of selected features and classification accuracy.

Experimental Results using the Iris Dataset

The experiment involves a comparative analysis of the LTV feature selection with several prominent feature selection methods to assess its ability to appropriately rank the Iris dataset's features. For this purpose, Relief-F [26], unsupervised feature saliency (UFSA) [27], Laplacian Score [28], unsupervised feature selection through fitness proportionate sharing clustering (UFSFPS) [29], locally linear embedding (LLE) [30] and radial basis function neural networks (RBFNN) [31] were used. Table 2 shows how the proposed and other methods ranked the features in the Iris dataset.

It can be observed that $\mathbf{f}_3$ and $\mathbf{f}_4$ as more significant features than $\mathbf{f}_1$ and $\mathbf{f}_2$, as depicted in Table 2. The LTV can choose a similar sequence of features based on their relevancy, just like those prominent methods. Therefore, it can be inferred that the LTV feature selection has the capability to effectively prioritise features according to their significance level.

Table 2. Ranking of the Iris dataset based on feature selection method.

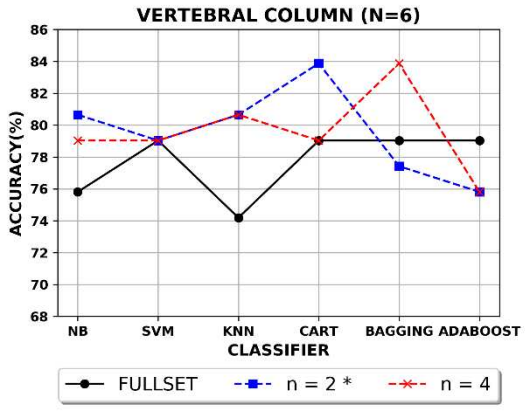
Method	Features Ranking
LTV	f_3, f_4, f_1, f_2
UFSFPS	f_3, f_4, f_1, f_2
Relief-F	f_4, f_3, f_1, f_2
LLE Score	f_3, f_4, f_1, f_2
UFSA	f_3, f_4, f_1, f_2
RBFNN	f_4, f_3, f_1, f_2
Laplacian Score	f_4, f_3, f_1, f_2

Experimental Results using Unknown Structure Datasets

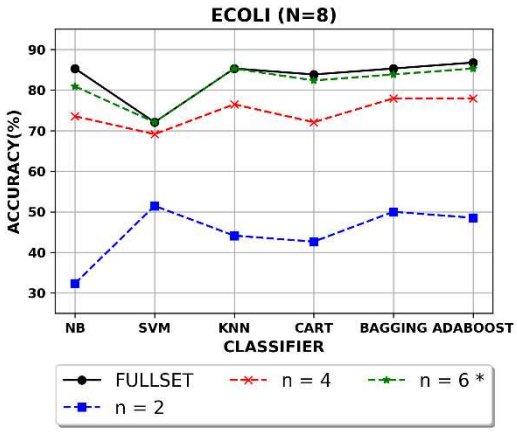
A comparison between the LTV feature selection and AmRMR method in terms of classification performance is presented next. This comparison is intended to see whether the selected feature subset can effectively rank features according to their relevancy and reprsent the full feature set.

Classification Performance Given by the Selected Feature Subsets

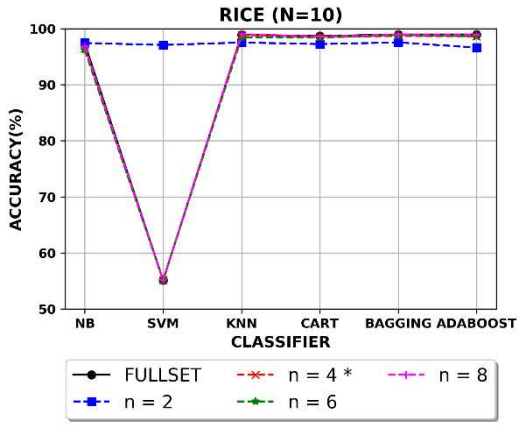
The performance of the feature subsets selected by the LTV was evaluated based on the first *n* features in the feature relevance ranking. The classification accuracy obtained by each selected subset and the full dataset was compared, while NB, SVM, KNN, CART, Bagging, and AdaBoost were employed as the classifiers in this evaluation. Figure 2 illustrates the subsets' ability to represent the full dataset.



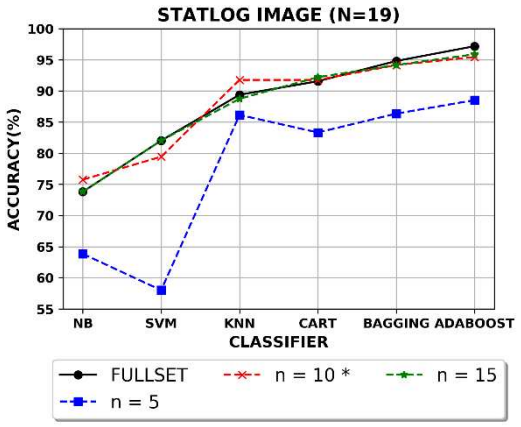
(a)



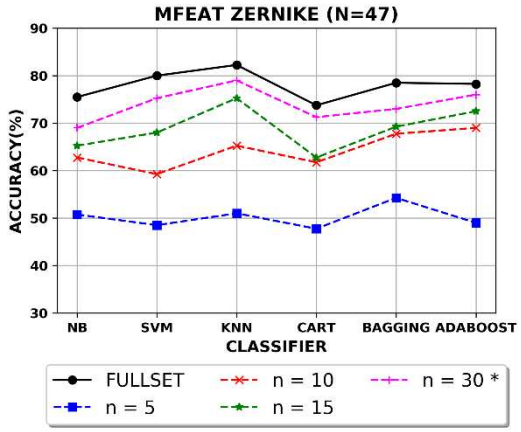
(b)


RICE (N=10)
ACCURACY(%)
CLASSIFIER
NB
SVM
KNN
CART
BAGGING
ADABOOST
FULLSET
n = 2
n = 4 *
n = 6
n = 8

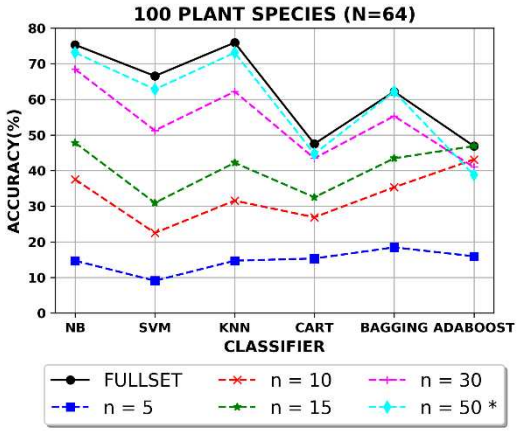
(c)


STATLOG IMAGE (N=19)
ACCURACY(%)
CLASSIFIER
NB
SVM
KNN
CART
BAGGING
ADABOOST
FULLSET
n = 5
n = 10 *
n = 15

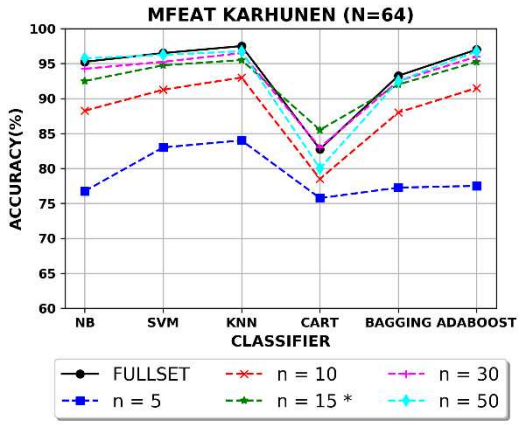
(d)


MFEAT ZERNIKE (N=47)
ACCURACY(%)
CLASSIFIER
NB
SVM
KNN
CART
BAGGING
ADABOOST
FULLSET
n = 5
n = 10
n = 15
n = 30 *

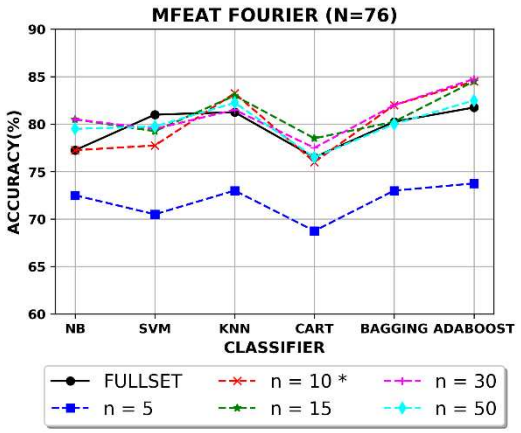
(e)


100 PLANT SPECIES (N=64)
ACCURACY(%)
CLASSIFIER
NB
SVM
KNN
CART
BAGGING
ADABOOST
FULLSET
n = 5
n = 10
n = 15
n = 30
n = 50 *

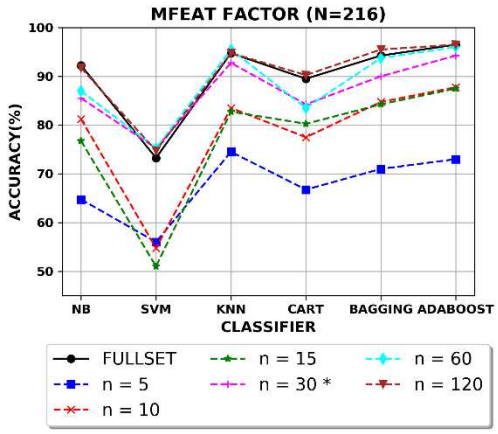
(f)


MFEAT KARHUNEN (N=64)
ACCURACY(%)
CLASSIFIER
NB
SVM
KNN
CART
BAGGING
ADABOOST
FULLSET
n = 5
n = 10
n = 15 *
n = 30
n = 50

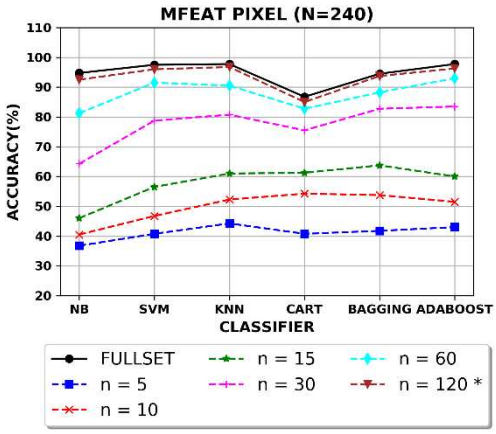
(g)


MFEAT FOURIER (N=76)
ACCURACY(%)
CLASSIFIER
NB
SVM
KNN
CART
BAGGING
ADABOOST
FULLSET
n = 5
n = 10 *
n = 15
n = 30
n = 50

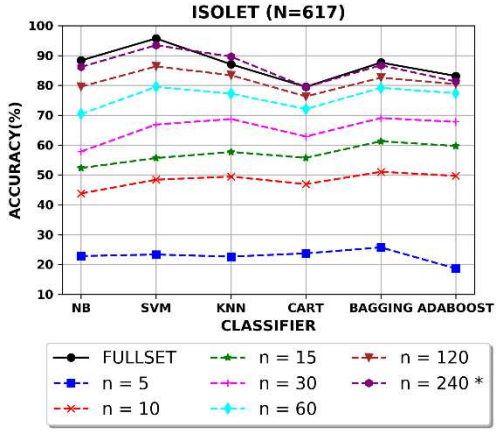
(h)



(i)



(j)



(k)

Figure 2. Classification accuracy obtained by feature subsets of n features and the full feature set.

Based on Figure 2, seven (Vertebral Column, Rice, Mfeat Karhunen, Mfeat Fourier, Mfeat Factor, Mfeat Pixel, and Isolet) out of eleven datasets achieved an accuracy level comparable to that of the full feature set while employing different classifiers by selecting either half or less than the full feature set. The LTV selected over 50% of the full feature set for four datasets. Six out of eight features were selected for the Ecoli dataset, 10 out of 19 features for the Statlog dataset, 30 out of 47 features for the Mfeat Zernike dataset, and 50 out of 64 features for the 100 Plant Species dataset. In general, it can be inferred that the feature subsets yield by the LTV feature selection are representative the full feature sets with a small feature subset size.

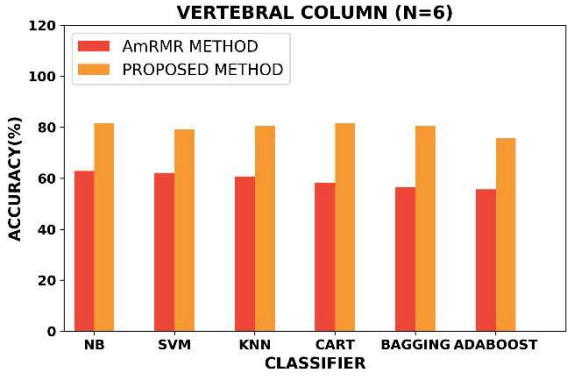
Table 3. The average classification accuracy achieved by different classifiers.

Classifier	**Average Classification Accuracy (%)**		**Difference in Accuracy (%)**
	Full feature set N	Feature subset n_{least}	
NB	84.63	83.11	1.52
SVM	79.89	78.00	1.89
KNN	87.66	86.44	1.22
CART	80.85	79.43	1.42
Bagging	86.24	84.83	1.41
AdaBoost	85.75	83.91	1.84

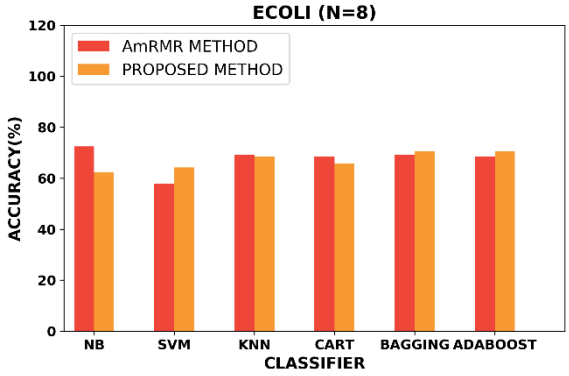
Table 3 displays the average classification accuracy obtained by different classifiers over a total of 11 datasets. The proposed LTV feature selection demonstrated slightly better performance accuracy when the results based on the KNN classifier were observed. Although the average classification accuracy yielded by the selected feature subset is lower than that of other classifiers, there is only 1% to 2% variation when compared to the accuracy provided by the full feature set. Thus, one could deduce that the proposed method can produce a satisfying feature subset that is representative of the full feature set since that those classifiers exhibit little difference in accuracy.

Comparison with the AmRMR Method

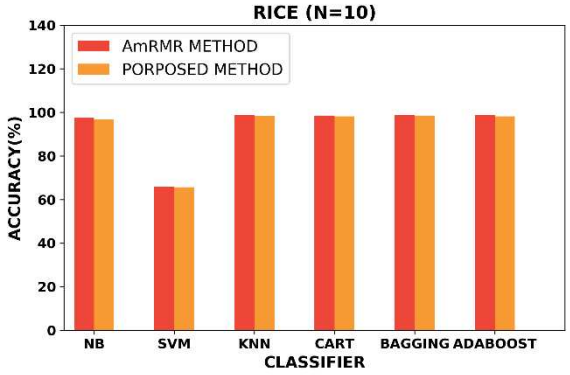
A comparative evaluation of the proposed LTV feature selection and the AmRMR method was conducted to assess the classification accuracy given by them. Figures 3 and 4 compare the results of the two methods. The *y*-axis of the graphs in both figures reflects the average classification accuracy based on the first *n* selected features, considering all test datasets.



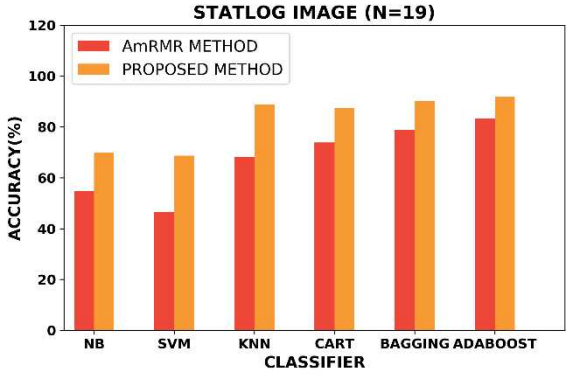
(a)



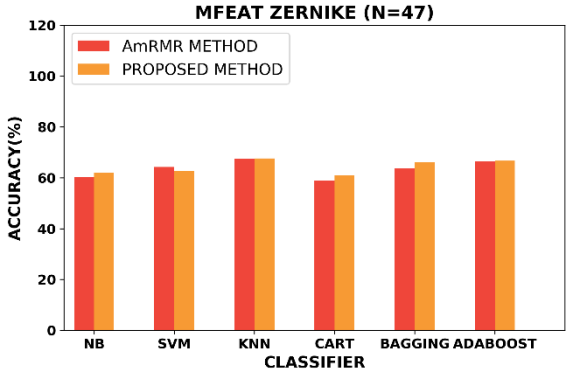
(b)



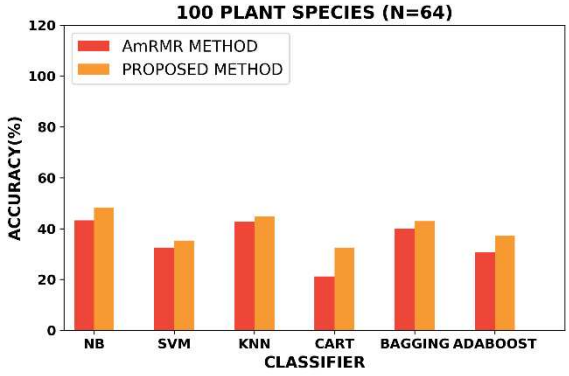
(c)



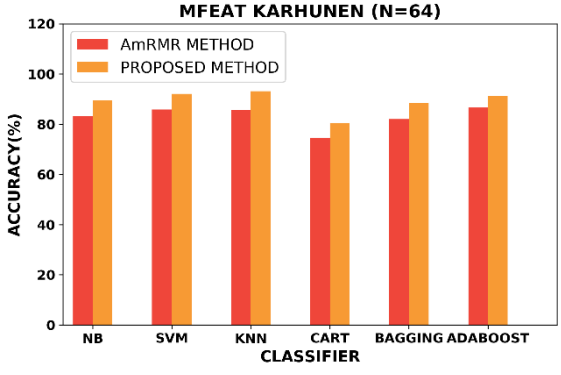
(d)

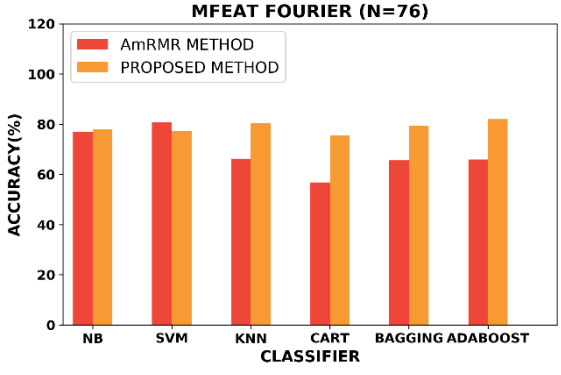


(e)

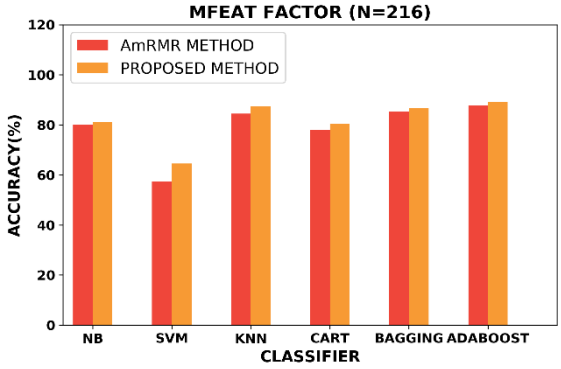


(f)

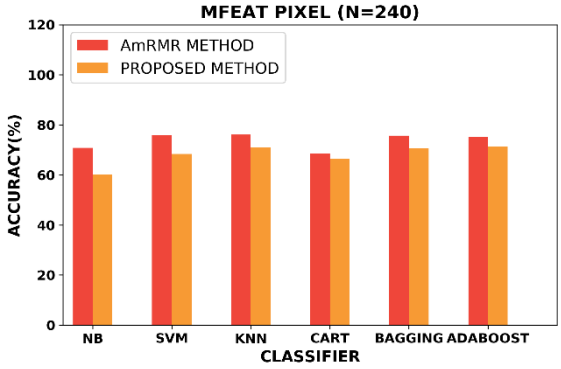




(g)

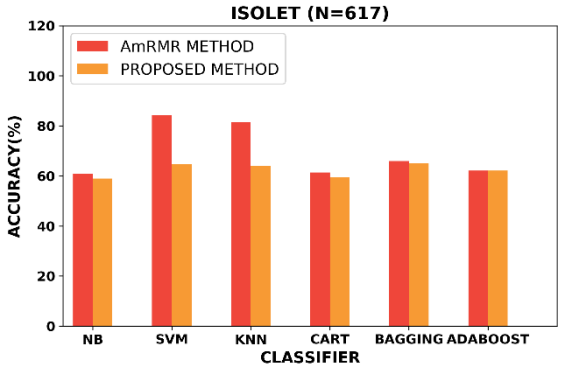


(h)



(i)

(j)



(k)

Figure 3. Average classification accuracy for the selected first *n* features.

Figure three shows that the Vertebral Column, Statlog Image, 100 Plant Species, Mfeat Karhunen, and Mfeat Factor datasets exhibit better performance when subjected to the LTV feature selection compared to the AmRMR method. Each classifier show comparable results when applied to the Rice dataset. On the other hand, it can be observed that the AmRMR method exhibited better performance for all classifiers when applied to the Mfeat Pixel dataset. When the LTV feature selection was used to the Mfeat Fourier dataset, it showed higher accuracy when compared to the AmRMR method, except when employed with the SVM classifier. When the NB and CART classifiers were observed, it was found that the LTV method gave lower classification accuracy than the AmRMR method on the Ecoli dataset. However, both methods give comparable performance when using the KNN classifier for this dataset. The SVM classificatoon accuracy was found to be lower on the Mfeat Zernike dataset when feature subset results from the LTV method were tested, while the KNN and AdaBoost classifiers achieved results of comparable accuracy. Both methods yielded comparable results on the Isolet dataset, with an exception for the SVM and KNN classification. When it comes to classification tasks, the LTV feature selection usually works better than the AmRMR method. This is because the generated feature subsets by the LTV method can give either better or similar classification accuracy, but not on the Mfeat Pixel data.

The results are further analysed based on Table 4. The goal is to see which method is better for classification tasks based on all six classifiers considered before.

Table 4. Average classification accuracy yielded by the LTV and AmRMR methods across different classifiers.

Classifiers	Classification accuracy (%)	
	LTV method	AmRMR method
NB	70.70	71.71
SVM	66.86	67.24
KNN	75.51	74.22
CART	69.66	64.53
Bagging	74.79	70.67
AdaBoost	74.62	70.85

Due to the highest classification accuracy shown by the KNN classifier in comparison to other classifiers, this classifier is considered the best fit for the LTV feature selection. According to the results presented in Table 4, it can be observed that most of the classifiers, precisely four out of six, exhibit a higher level of compatibility with the LTV method. Although the AmRMR method yields better accuracy when applied to the NB and SVM classifiers, the LTV method can

still provide dependable performance since the difference in accuracy between the two methods is just marginal. Hence, the LTV method generally beats the AmRMR method.

CONCLUSION

This paper presents the LTV feature selection, a supervised feature selection that utilises a correlation-based measure to determine the relevance of features. The findings suggest that the LTV feature selection shows adequate reliability in representing the full feature set using only a small feature subset size. While the method may not consistently yield the optimal feature subset, it demonstrates sufficient capability for minimising the data's dimensionality and generating highly representative data. When evaluated with all six classifiers, the LTV could also produce results with excellent accuracy, particularly with the KNN classifier. In addition, the LTV was superior to the AmRMR method in the context of the classification task.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to the International Islamic University Malaysia (IIUM), Universiti Malaysia Pahang (UMP) and the Universiti Teknologi MARA (UiTM) for financial support provided through the IIUM-UMP-UiTM Sustainable Research Collaboration Grant 2020 (Vote Number: RDU200722).

REFERENCES

[1] Idrees SM, Alam MA, Agarwal P. A study of big data and its challenges. Int J Inf Technol 2019;11:841–6.
[2] García S, Luengo J, Herrera F. Data Preprocessing in Data Mining. vol. 72. Cham: Springer International Publishing; 2015.
[3] Han J, Kamber M, Pei J. Data preprocessing. Data Min. Concepts Tech., Elsevier; 2012, p. 83–124.
[4] Abdel Maksoud EA, Barakat S, Elmogy M. Medical Images Analysis Based on Multilabel Classification. Elsevier Inc.; 2019.
[5] Saberi-Movahed F, Eftekhari M, Mohtashami M. Supervised feature selection by constituting a basis for the original space of features and matrix factorization. Int J Mach Learn Cybern 2020;11:1405–21.
[6] Solorio-Fernández S, Martínez-Trinidad JF, Carrasco-Ochoa JA. A Supervised Filter Feature Selection method for mixed data based on Spectral Feature Selection and Information-theory redundancy analysis. Pattern Recognit Lett 2020;138:321–8.
[7] Gan J, Wen G, Yu H, Zheng W, Lei C. Supervised feature selection by self-paced learning regression. Pattern Recognit Lett 2020;132:30–7.
[8] Lim H, Kim DW. Pairwise dependence-based unsupervised feature selection. Pattern Recognit 2021;111:107663.
[9] Zheng W, Zhu X, Wen G, Zhu Y, Yu H, Gan J. Unsupervised feature selection by self-paced learning regularization. Pattern Recognit Lett 2020;132:4–11.
[10] Chen X, Yuan G, Nie F, Ming Z. Semi-Supervised Feature Selection via Sparse Rescaled Linear Square Regression. IEEE Trans Knowl Data Eng 2020;32:165–76.
[11] Shi C, Gu Z, Duan C, Tian Q. Multi-view adaptive semi-supervised feature selection with the self-paced learning. Signal Processing 2020;168:107332.
[12] Shukla AK, Pippal SK, Gupta S, Ramachandra Reddy B, Tripathi D. Knowledge discovery in medical and biological datasets by integration of Relief-F and correlation feature selection techniques. J Intell Fuzzy Syst 2020;38:6637–48..
[13] Rani P, Kumar R, Jain A. A Hybrid Approach for Feature Selection Based on Correlation Feature Selection and Genetic Algorithm. Int J Softw Innov 2022;10:1–17.
[14] Al-juboori AM, Alsaeedi AH, Nuiaa RR, Alyasseri ZAA, Sani NS, Hadi SM, et al. A Hybrid Cracked Tiers Detection System Based on Adaptive Correlation Features Selection and Deep Belief Neural Networks. Symmetry (Basel) 2023;15:1–14.
[15] Heredia-Márquez A, Guzmán-Arenas A, Martínez-Luna GL. Feature Selection Ordered by Correlation - FSOC. Comput y Sist 2023;27:33–51.
[16] Senawi A, Wei HL, Billings SA. A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking. 2017.
[17] Soch J. The Book of Statistical Proofs 2022. https://statproofbook.github.io/P/var-tot.
[18] Mustapa NA, Senawi A, Liang CZ. Feature Selection Using Law of Total Variance with Fast Correlation-Based Filter. 2023 IEEE 8th Int. Conf. Softw. Eng. Comput. Syst., IEEE; 2023, p. 35–40.
[19] Schober P, Schwarte LA. Correlation coefficients: Appropriate use and interpretation. Anesth Analg 2018;126:1763–8.
[20] Jo I, Lee S, Oh S. Improved Measures of Redundancy and Relevance for mRMR Feature Selection. Computers 2019;8:42.
[21] Dua D, Graff C. UCI Machine Learning Repository. Univ California, Irvine, Sch Inf Comput Sci 2019. http://archive.ics.uci.edu/ml.
[22] Settouti N, Bechar MEA, Chikh MA. Statistical comparisons of the top 10 algorithms in data mining for classification task. Int J Interact Multimed Artif Intell 2016:46–51.
[23] Guo G, Wang H, Bell D, Bi Y, Greer K. KNN Model-Based Approach in Classification. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 2888, 2003, p. 986–96.
[24] Zhang S, Li X, Zong M, Zhu X, Cheng D. Learning k for kNN Classification. ACM Trans Intell Syst Technol 2017;8:1–19.
[25] Zhongguo Y, Hongqi L, Ali S, Yile A. Choosing classification algorithms and its optimum parameters based on data set characteristics. J Comput 2017;28:26–38.
[26] Liu Z, Wen T, Sun W, Zhang Q. Semi-Supervised Self-Training Feature Weighted Clustering Decision Tree and Random Forest. IEEE Access 2020;8:128337–48.
[27] Swathi BV. Performance Analysis of a Gaussian Mixture based Feature Selection Algorithm. IJRITC 2017;5:150–5.
[28] Yankee TN. Rank Aggregation of Feature Scoring Methods for Unsupervised Learning. 2017.
[29] Yan X, Homaifar A, Awogbami G, Girma A. Unsupervised Feature Selection through Fitness Proportionate Sharing Clustering. Proc - 2018 IEEE Int Conf Syst Man, Cybern SMC 2018 2019:1355–60.

[30] Yao C, Liu YF, Jiang B, Hen J, Han J. LLE Score: A New Filter-Based Unsupervised Feature Selection Method Based on Nonlinear Manifold Embedding and Its Application to Image Recognition. IEEE Trans Image Process 2017;26:5257–69.
[31] Zeng X, Zhen Z, He J, Han L. A feature selection approach based on sensitivity of RBFNNs. Neurocomputing 2018;275:2200–8.