Ann. Data. Sci. (2017) 4(4):483–502
DOI 10.1007/s40745-017-0116-1

A Feature Selection Method Based on Ranked Vector Scores of Features for Classification

Firuz Kamalov[1] · **Fadi Thabtah**[2]

Received: 13 January 2017 / Revised: 24 May 2017 / Accepted: 13 July 2017 / Published online: 29 July 2017

Abstract One of the major aspects of any classification process is selecting the relevant set of features to be used in a classification algorithm. This initial step in data analysis is called the feature selection process. Disposing of the irrelevant features from the dataset will reduce the complexity of the classification task and will increase the robustness of the decision rules when applied on the test set. This paper proposes a new filtering method that combines and normalizes the scores of three major feature selection methods: information gain, chi-squared statistic and inter-correlation. Our method utilizes the strengths of each of the aforementioned methods to maximum advantage while avoiding their drawbacks—especially the disparity of the results produced by these methods. Our filtering method stabilizes each variable score and gives it the true rank among the input data's available variables. Hence it maximizes the stability in the variables' scores without losing the overall accuracy of the predictive model. A number of experiments on different datasets from various domains have shown that features chosen by the proposed method are highly predictive when compared with features selected by other existing filtering methods. The evaluation of the filtering phase was conducted via thorough experimentations using a number of predictive classification algorithms in addition to statistical analysis of the filtering methods' scores.

✉ Firuz Kamalov
firuz@cud.ac.ae

Fadi Thabtah
f.thabtah2@hud.ac.uk

1 Canadian University of Dubai, Dubai, UAE

2 University of Huddersfield, Huddersfield, UK

Keywords Classification accuracy · Data mining · Dimensionality reduction · Feature selection · Predictive models · Ranking of features

1 Introduction

Data mining plays an important role in uncovering hidden information within data. It is used many fields including business, medicine, security and others [5,8,11,15,21,27]. One of the primary elements that influence the construction of predictive models in data mining is the choice of variables to be utilised during the construction of the model. In practice, this can be problematic when there are large numbers of variables in the dataset. Therefore, it is necessary to filter out the extraneous variables to reduce the dimensionality of the features space and leave only the relevant features to be used in the classifier. In fact, the feature selection phase has been shown to have an impact on the quality of the outcome in a wide range of applications including text mining [20], website security [2], and others [1,3,4,6,7,10,13,16,22,24].

The aim of a machine learning classifier is to construct a set of rules that would predict the correct output based on the input. The input variables are called features and the output variable is referred to as the target class. The set of rules that are used to produce the value of the target class based on the input values of the features is called the predictive model. The predictive model is built based on a training data where one tries to discern the relationship between the features and the target class. Training data often contains a lot of noise in the form of irrelevant features. It is therefore important to be able to filter out the redundant variables to improve the performance of the predictive model. The reduction in the dimensionality of the feature space allows researchers to better understand the predictive model and the nature of the relationship between the features and the target class. Finally, a fewer number of features reduces the computational load on the processing system.

Typically, there are two common feature selection approaches utilized in the literature: filtering and wrapping methods [14]. In filtering methods such as Information Gain [17], each variable in the training dataset is assessed by computing its relevancy with the target attribute (class label). Any variable that has a gain larger than a predefined threshold is said to be relevant and therefore is kept for further data processing. Whereas any variable that fails to pass the predefined threshold gets removed. The variable quality is measured based on the filtering method used. For instance, IG measures the variable significance by calculating the reduction in entropy of the target variable when the information about the feature variable is known. The feature variable that results in the greatest reduction of entropy in the target variable is chosen. Further details of this approach are given in Sect. 2.

The second approach to feature selection are the wrapper methods such as the Feature Elimination algorithm [25] which utilize the outcome of the data mining technique to assess the variables' quality in the input dataset. In wrapper methods, a number of different combinations of the available variables are tested and contrasted to other combinations. These methods consider the variable selection as a search problem. Variables are chosen in wrapper methods based on the performance results

of the predictive algorithms and consequently methods which fall under this type of feature selection, have been criticized for being slow [14].

One of the major challenges in feature selection is the inconsistency in terms of the selected features by various methods. For instance, if we run two common filtering methods IG [17], or Chi Square [13] on the "Labour" and "Hepatitis" datasets from the University of Irvine data collection [12], we will end up with different chosen variables. In particular, IG selects 14 and 17 variables from "Labour" and "Hepatitis" datasets respectively using the predefined threshold of 0.01. On the other hand, Chi Square selects 3 and 9 features respectively from the same datasets using a predefined threshold of 10.83 [25]. The results also may vary more significantly if the user has decided to use different thresholds other than the default ones used in both filtering methods. This example, although limited, illustrates high discrepancies in results obtained by applying different feature selection methods. Hence a comprehensive method that may reduce this discrepancy is needed.

We believe that combining feature scores using filtering methods can reduce variations in the current filtering methods' results and provide higher confidence in the scores assigned to variables. This may be seen as a unified way of having multiple filtering methods contributing to a cumulative score per variable which may stabilize that score and give it the true weight and rank among other variables. Hence, we introduce a new filtering method for feature selection that reduces the instability of the new variable score without losing in the overall accuracy of the predictive model. The proposed filtering method was influenced by the portfolio diversification idea in finance [5] in which the investor can own a basket of unrelated securities to reduce the investment risk. In other words, an investor may sustain the same level of return while lowering his portfolio's risk by merging uncorrelated assets. Thus in the context of predictive models in data mining, combining different scoring methods should stabilise the classification accuracy across various different datasets while maintaining the overall average classification accuracy.

Our proposed method consists of two steps. First, we evaluate, normalize, and merge the variables' scores from the two existing filtering methods (IG, Chi Square) to come up with a unified score vector for each variable. We then compute the norm of the score vector and use it to discriminate among the feature variables. In particular, variables with larger magnitude will have higher score and hence will be more likely to be selected. Unlike some other existing methods that combine variables' scores from different filtering methods such as AND and OR, our approach yields a new true metric on the space of all pairs of scores. This allows for a mathematical structure to analyze the space of combined scores. More details on how the magnitude is computed and other mathematical formulas are given in Sect. 3. The second step of our selection algorithm is to use the score vectors computed for each variable to filter the subset of variables chosen via the CSF method. Recall that the CSF method attempts to produce a subset of features that have a high correlation with the target class while small correlation among the features. The primary research question in this paper is "whether the new filtering method will further minimize the search space of variables without significantly hindering the predictive models accuracy?"

Another advantage of a unified variable metric is the fact that fewer numbers of variables will be chosen. This will happen because only the variables with limited

correlations among each other are kept and thus fewer computing resources will be demanded by the predictive model during the mining process. Moreover, the classification systems derived will possibly contain a more concise set of knowledge especially in rule based predictive models. Therefore, managers and decision makers have the ability to exert more control and understand the content of these classification models. This can be obvious in application domains that necessitate fewer but effective knowledge base such as medical diagnoses and cyber security. The predictive models selected to measure the proposed filtering method effectiveness are eDRI [19], C4.5 [18] and PART [25]. The choice of these predictive algorithms is based on the fact that they generate If-Then models that can easily be interpreted by users as well as their widespread use in business domains.

Common filtering methods used in this paper and related works are reviewed in Sect. 2. Section 3 is devoted to the proposed filtering method and Sect. 4 contains experiments and results analysis. The conclusions and further research are given in Sect. 5.

2 Feature Selection Methods: A Quick Review on Combining Features

Information gain (IG) is one of the most popular feature selection methods. In [4], the authors constructed an algorithm based on IG for selecting relevant features from intrusion detection datasets. In [15] used IG as part of a three step algorithm to find the optimal subset of features to increase classification accuracy and scalability in credit risk assessment. In [10], IG was employed to propose a greedy feature selection method.

This method works by ranking features according to their IG score. The IG score of a feature is obtained by calculating the mutual information between the class label and the feature.

$$I\,(C, A) = H\,(C) - H\,(C|A) \tag{1}$$

where C is the class variable, A is the attribute variable, and $H()$ is the entropy. Features with higher IG scores are ranked above the features with lower scores. The IG method was first proposed by Quinlan and was implemented by him in his ID3 decision tree algorithm. Thereafter, IG has become a major evaluation tool in feature selection.

Another important method in feature selection is the chi squared method (CHI). It is a widely used metric in machine learning for evaluating the goodness of an attribute [13]. In [6], the authors evaluated various feature selection methods and showed that CHI performed very well under the "stability" criteria. In addition, in [1], Support Vector Machine (SVM) classifiers are used for sentiment analysis with several univariate and multivariate methods for feature selection. Moreover, [8] utilized CHI in the filtering process as part of their method for sentiment analysis.

This method is based on calculating the chi squared statistic between the class variable and the feature variable in the data. To compute the CHI score X is the number of times feature a and class c occur together, Y is the number of times feature a occurs without class c, W is the number of times class c occurs without feature a,

Z is the number of times neither a or c occur, and N is the total size of the training set. Then the CHI score is given by

$$CHI\,(a,c) = \frac{N \times (XZ - YW)}{(X+W) \times (Y+Z) \times (X+Y) \times (W+Z)} \tag{2}$$

The chi squared test has long been used by statisticians to gauge the degree of independency between a pair of categorical variables. Thus it was natural to adopt the CHI score as a way of determining a feature's relevancy with respect to the class in the context of data mining.

In [21] the authors proposed a new method for combining the IG and CHI scores in their recent published paper. Since the IG score and the CHI score are highly correlated it seems natural to combine them into one comprehensive score. By taking the (IG, CHI) score of a feature as a vector the authors proposed a new score by computing the magnitude of this vector

$$|v_a| = \sqrt{(IG_a)^2 + (CHI_a)^2} \tag{3}$$

For simplicity we will refer to this score as simply the V-score. It was shown that the V-score improves the accuracy of the classification algorithms when applied to phishing data.

Information gain, chi squared and V-score are each good measures of relevancy of an individual feature with respect to the class. However, in practice we often need to choose a subset of features to use in the classification algorithm. A simple approach to forming the optimal subset of features would be to select features based on their scores. However, this approach would ignore any interactions between the features. For instance, if two features have a high V-score with respect to one another then it would be redundant to choose both features in the optimal feature subset.

One way to account for interactions between the features is to use the Correlation Feature Selection (CFS) method [9]. The CFS method evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred. The CSF criterion is computed by the following method

$$CFS = max_{S_k} \frac{r_{yx_1} + r_{yx_2} \cdots + r_{yx_k}}{\sqrt{k + 2\left(r_{x_1x_2} + \cdots r_{x_kx_1}\right)}} \tag{4}$$

The main drawback of the CFS method is that it often produces subsets of features that contain features with relatively low IG and V-scores. Therefore, we propose to combine the results of the CFS method together with the V-scores.

Several authors have used various methods in the past to combine the information from various feature selection methods into one score [3,23,24,26]. Thubaity et al. examined the effects of combining five feature selection methods, namely CHI, IG, GSS, NGL, and RS on Arabic text classification. Two approaches in combination

were used, intersection (AND) and union (OR). The experiments showed a minor improvement in classification accuracy for combining two and three feature selection methods. No improvement in accuracy was observed when more than three features were combined.

In [23] the authors considered combining multiple feature selection methods to identify more representative variables for predicting stock prices. Search methods, used are Principal Component Analysis (PCA), Genetic Algorithm (GA) and decision trees (CART). The combination methods to filter out undesirable variables were based on union, intersection, and multi-intersection strategies. It was shown that the intersection between PCA and GA and the multi-intersection of PCA, GA, and CART performed the best. In addition, these two combined feature selection methods filter out nearly 80% of unrepresentative variables.

Rogati and Yang examined the benefits of combining uncorrelated methods in the context of text classification. It was shown that in some cases the combined score enhanced the performance of the combined selection method. The combination of two methods was performed by normalizing the scores for each word and taking the maximum of the two scores, essentially performing OR with equal weights.

Thabtah and Kamalov proposed a method that merges the IG and CHI squared scores of a feature into a single score. They showed that this approach reduced variability of the ranking of features without affecting the accuracy of a classifier. In particular, the authors tested the new scoring method on phishing data and showed its advantages compared to using IG or CHI squared alone. In particular, after applying data mining algorithms on the features identified by the new method, IG and CHI, the accuracy of the classifiers derived from the set of features that was chosen by their method outperformed those of CHI and IG. In addition, using a combined score allows for more consistent results than using a single score.

In [16], the authors investigated different feature selection methods on the problem of sentiment analysis. They proposed a filtering method called Expansion Ranking (ER) that takes of a query results in information retrieval and assigns weights to each feature in the sentiment dataset. Experimental results on four Turkish product review datasets and using Naïve Bayes classifier revealed that ER selects features that when processed yields higher accuracy than CHI filtering method at least on the Turkish review datasets.

Feature pre-processing has been investigated by [11] to determine the influential features in medical images. The authors have concentrated on two methods, i.e. Genetic Search and greedy stepwise. They concluded that integrating Genetic search with greedy stepwise methods might yield smaller effective features.

Zhou et al. [28] proposed a filtering method based on interclass and intraclass relative frequencies of attribute values in the dataset. The authors identified three primary factors related to each feature during pre-processing, i.e. term frequency, interclass relative frequency and intraclass relative frequency. These frequencies of terms are used to assign scores for each feature during pre-processing and then features' scores above a predefined threshold are chosen. Experiments against textual dataset showed that the proposed filtering method scales well if compared with simple filtering methods such as Document Frequency (DF).

Yousef et al. [27] evaluated a number of feature selection methods on MicroRNAs dataset with more than 700 features aiming to improve classification accuracy. MicroRNAs are RNA sequences concerned with posttranscriptional gene regulation. After evaluating a number of feature selection methods against MicroRNAs dataset, the authors reported that clustering the features might improve the performance of the machine learning method especially when irrelevant features are removed from each cluster.

3 The Proposed Filtering Method

One way to ease the data mining process when the input dataset is high dimensional is to carefully select the relevant features [9]. This usually influences the predictive performance of the classification algorithm. A researcher often starts with data that contains multiple features that are meant to help identify the correct class labels. Naturally, some features would be more relevant to the class label than others while some can be altogether irrelevant. Therefore, it is important to select the features that will be most helpful in identifying the correct class labels in the data at a preliminary stage.

IG and CHI are widely used metrics in feature ranking. Researchers often compute IG and CHI scores of the features and then decide which features are to be selected based on some pre-determined criteria. However, the features ranking produced by each method may not match. Features ranked high by the IG criterion may not necessarily be ranked high by the CHI criterion. In this case, selecting the right features becomes a not so straightforward task. In [6], it was shown that the goodness rate of IG and CHI are highly uncorrelated. Therefore, one cannot expect the same features to be selected by the IG scoring as by the CHI scoring.

One way to address possible contradictions between the IG ranking and the CHI ranking is to combine the IG and CHI scores into one single score. Since IG and CHI scores are uncorrelated then combining the two scores would produce a more stable score. This approach is akin to the idea of portfolio diversification in finance where investors combine uncorrelated assets to reduce the overall volatility of the portfolio. In the context of feature selection combining the IG and CHI scores would stabilize the ranking of features compared with single metric approach.

We propose to create a vector of scores based on the IG and CHI scores and compute a "V-score" as the magnitude of the vector. In our filtering approach, we rely on the V-score together with the CSF method to produce a new, more robust criterion to feature selection. Before proceeding further with our discussion we give the necessary background on V-score as it is a key component of our new method.

The IG and CHI scores have very different values. Therefore, to combine them we first need to normalize both scores to make them comparable. So let IG_{max} denote the maximum IG score among all the available features then define the normalized score of the ath attribute by

$$\overline{IG_a} = \frac{IG_a}{IG_{max}} \tag{5}$$

Likewise we normalize the CHI scores by

$$\overline{CHI_a} = \frac{CHI_a}{CHI_{max}} \quad (6) \tag{6}$$

We next define the *score vector* of feature a to be

$$v_a = \begin{pmatrix} IG_a \\ CHI_a \end{pmatrix} \tag{7}$$

The score vector thus contains information about both IG and CHI scores. Recall that the magnitude of a vector is given by the square root of the sum of squares of its coordinates. Therefore, the magnitude of the score vector can be used as a scalar metric of the vector

$$|v_a| = \sqrt{(IG_a)^2 + (CHI_a)^2} \tag{8}$$

The magnitude of the score vector can be used to compare feature to one another. Features with greater value of $|v_a|$ will be ranked higher. Unlike other ways of combining scores from different methods such as AND and OR, our approach yields a true metric on the space of all pairs of scores. This allows for a mathematical structure for analyzing the space of combined scores.

As we mentioned in Sect. 2, the main drawback of the CFS filtering method is that it often produces subsets of features that contain features with relatively low IG and V-scores. We attempt to correct this problem by combining the results of the CFS method together with the V-scores. In particular, we propose a two-step process for creating the optimal subset of features used for classification. First, given the initial set of features, we apply the CSF method to obtain the correlation-based subset. Then we further reduce the correlation-based subset by selecting the features with relatively high V-scores. In particular, we select the features with the V-scores of 50% or above of the maximum V-score in the CSF subset. For instance, if the maximum V-score in the set is 1.2 then we choose the features with a V-score of 0.6 or above.

The key to our approach is to reduce the number of features in the CFS subset by removing those features with relatively low V-scores. In order to decide on the cut off V-score, we initially applied a threshold limit of 5% of the top V-score of the feature set. For instance, if the highest V-score in the CFS subset was 0.9, then we would eliminate any feature with a V-score of 0.045 or less. We also experimented with the threshold of 10%. However, in both cases we realized that very few features in the CSF feature subset did not meet the threshold score, i.e. the size of the feature subset did not change by much. Therefore, we applied a higher threshold of 50%. So for instance, if the highest V-score in the CFS subset was 0.9, then we would eliminate any feature with a V-score of 0.45 or less. This cut-off point made a substantial difference in the size of the CFS subset and provided a real gain in terms of feature selection.

To determine the cut-off V-score to qualify as a feature for selection to the optimal set, we must be aware of the trade-off between the number of features and the accuracy of a classifier. If we set the benchmark too high (e.g. 70%) then we would substantially

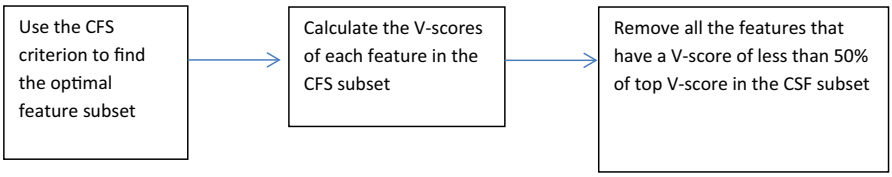

Fig. 1 The proposed feature method selection primary processes

decrease the number of features used for classification which would be advantageous. However, having too few features may have a substantial negative impact on the accuracy of the classifiers. On the other hand, if we increase the number of features used for classification then we would almost certainly increase the fitness of the classifier even if only by a very small margin. Taken to its extreme, this approach will lead to including all the features in the optimal features subset. However, this would result in several problems including overfitting so that the classifier will not generalize very well to other data. Therefore, the cut-off benchmark depends on individual preferences regarding the number of features and accuracy of the feature subset and the corresponding generalization and overfitting trade-offs.

Combining the CFS method with our V-score allows us to take advantage of both ranking techniques. In fact, a more proper statement would be to say that our approach to combining the two methods reduces the disadvantages of each method. The V-score has the drawback of not considering possible interactions between features. It is possible with features with low individual V-scores to perform very well as a joint set. On the other hand the CSF method tends to downplay the individual performances of features. So by taking the CSF feature subset and then analyzing the individual performance of each feature of the subset we are able to reduce the size of the feature subset without affecting the accuracy of the subset. Figure 1 below illustrates the process of feature selection for our method:

We summarize the main advantages of the proposed filtering method below:

- Produces a smaller feature set than other known filtering methods, i.e. IG, Chi and CSF.
- Takes into account interactions among features which are ignored by the IG and Chi squared methods
- Provides a more stable criterion
- Reduce the discrepancies of the variables' scores by combining multiple scores from different methods
- Contributes in the reduction of the classifiers at least for rule-based classification predictive models without minimising the classification accuracy

4 Data and Experimental Analysis

4.1 Experimental Setting

In this section we discuss the setup of the experiment that we used to analyze the performance of our newly proposed method for feature selection. We will describe

the data used in the experiment, the selection methods used to benchmark our own method, and the classification models used to analyze the performance of the selection methods. We aim to answer the following questions in line with the research question of Sect. 1:

(1) Does the proposed feature selection method produce fewer selected features than the existing methods without affecting the accuracy of the classification model?
(2) Will the classifiers generated by the new filtering method contain fewer rules hence allowing decision makers to better understand the nature of the relationship between the features and the target class?
(3) Will the new score and rank per variable proposed by our filtering method help in identifying effective variables and hence put them in higher ranks so they have better chance of being chosen for data processing?

All the experimental runs were conducted using an open source machine learning tool called WEKA [25]. WEKA is a Java platform that contains implementation of various different machine learning algorithms including filtering methods. This tool was developed at the University of Waikato in New Zealand. The experiments have been performed on a computing machine with a 2.0 Ghz processor.

For the predictive models experiments (Sect. 4.3), we used a tenfold cross-validation method to generate the classifiers. Typically, tenfold cross-validation is employed during the process of deriving classifiers to reduce overfitting. In cross-validation, the input data is divided into 10 partitions in which 9 partitions are exploited to build the classifiers and the remaining partition is used to test the predictive power of the classifier. Normally, this process is repeated ten times to generate an average accuracy of the classifier.

4.2 Datasets

We used several datasets from the UCI data repository [12] to measure the performance of the proposed filtering method. The choice of the datasets is based on different factors including application domain, attribute types, dimensionality size and the number of data examples. To achieve reliable results we tried to select datasets that have continuous and discrete variables, both large and small in terms of number of examples and dimensionality, and which belong to different. We also selected datasets with noise such as missing values in order to reflect the reality of applications on the overall performance of our filtering method. The datasets characteristics are depicted in Table 1.

4.3 Filtering and Predictive Methods Used

To benchmark the new filtering method and its effect on predictive models, we have chosen three existing filtering methods for comparison. These methods are CFS, IG and CHI. They have been selected because of the following reasons:

(a) The proposed method is based on IG, CHI and CFS so it is natural to consider these methods.

Table 1 The datasets features

Dataset	Size	# of variables	# of classes	Missing values	Continuous variable(s)	Application
Vote	435	17	2	Yes	No	Politics
Glass	214	10	7	No	Yes	Criminology
Diabetes_Pima	768	9	2	No	Yes	Medical
Breast-cancer	286	10	2	Yes	Yes	Medical
german_credit	1000	21	2	No	Yes	Finance
Labor	57	17	2	Yes	Yes	Business
Cleve	303	12	2	Yes	No	Medical
Cylinder-bands	540	40	2	Yes	Yes	Industrial
Hepatitis	155	20	2	Yes	Yes	Medical
Hypothyroid	3772	30	4	Yes	Yes	Medical
Ionosphere	351	35	2	No	Yes	Physics
Mushroom	8124	28	2	Yes	No	Agriculture
Segment	2310	20	7	No	Yes	Image segmentation
Lung-cancer	32	57	3	Yes	No	Medical
Arrhythmia	452	280	16	Yes	Yes	Medical

(b) CFS is a competitive filtering method that usually reduces the number of picked features when compared to CHI and IG. So it will be ideal to compare our method with it.
(c) IG, CHI and CFS have been employed successfully in preprocessing a wide range of application datasets therefore their results have already been evaluated and generalized by previous scholars.

The predefined thresholds for IG and CHI were set to 0.01 and 10.83 respectively similar to [25] and [13]. Any variable with a score above these minimum thresholds will be chosen and otherwise the variable will be ignored.

A number of predictive models from data mining have been utilized to evaluate the performance of the filtering methods on classification systems. In particular, a recent dynamic learning associative classification called eDRI [19], a decision tree algorithm called C4.5 [18] and a rule induction algorithm based on partial decision tree called PART [7] have been used. Our selection of these predictive algorithms is based on the fact that they adopt different training procedures to derive the classification models. Additionally, the three of them produce classifiers with human interpretable rules.

The eDRI algorithm employs association rule to discover the rules based on predefined thresholds; minimum support and minimum confidence. This algorithm finds all possible relationships between each variable and the target class and discards those with frequencies less than the minimum support. Then it invokes a rule evaluation procedure that checks each rule on the training dataset to only keep rules that have correctly classified training data. All other rules, which are unable to classify training data, are discarded. On the other hand, C4.5 is a learning algorithm that uses Entropy; a dissimilarity measure to build decision trees. C4.5 always looks for variables that have high information gain in discriminating the target class and removes any variable with minimal information gain. When the tree is constructed, each path from the root to any leave is converted into a rule. Finally, PART algorithm adopts information theory and greedy learning to build partial decision trees and then transforms them into rule based classifiers.

4.4 Analysis of the Results

Table 2 shows the number of features chosen by the benchmark methods based on the datasets considered. It is clear that our proposed filtering method was able to substantially reduce the search space of the selected features when compared with the results produced by IG, CHI and CFS. In fact, our method consistently selected fewer variables in each of the datasets considered. For example, in the "Vote" and "Mushroom" datasets, our filtering method has chosen 2 and 1 variables respectively whereas CHI, IG and CFS have derived (14,13,4) and (21,20,4) respectively from the same datasets.

The overall average number of features derived from the datasets is depicted in Fig. 2.

This figure reveals that our filtering method significantly reduces the number of selected features in all datasets. In particular, the average percentage differences in the search space reduction of variables between our method and (CHI, IG, CFS) are 44.81,

Table 2 Number of variables chosen per dataset by the filtering methods

Dataset	# of variables	CFS	IG	CHI	New filtering method
Vote	17	4	13	14	2
Glass	10	7	8	8	6
Diabetes_Pima	9	4	8	8	1
Breast-cancer	10	5	6	5	4
german_credit	21	3	10	11	2
Cleve	12	6	10	10	6
Cylinder-bands	40	6	21	20	2
Hepatitis	20	10	17	9	4
Hypothyroid	30	5	6	11	2
Ionosphere	35	14	34	34	12
Mushroom	28	4	20	21	1
Segment	20	8	16	16	6
Lung-cancer	57	10	52	10	8
Arrhythmia	280	26	139	145	15

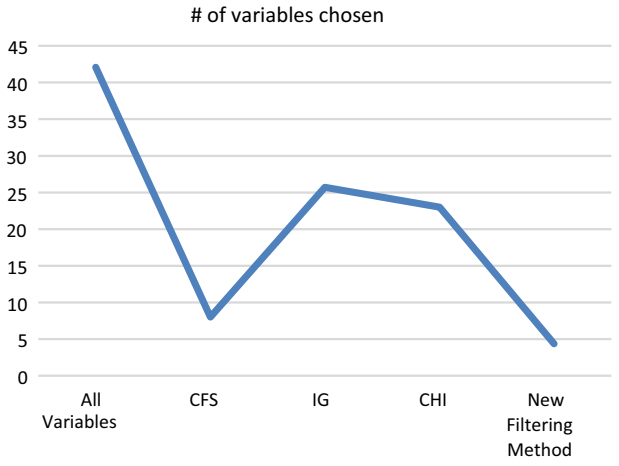

Fig. 2 Average number of features chosen from all datasets for each filtering method

50.76 and 8.66% respectively. These reductions are attributed to the new proposed score vector which takes into account the correlations among different combinations of variables per dataset and ensures only minimal variable-variable correlations and maximum variable-class correlations remain for data processing. This allows us to reduce the number of features used in classification with little effect on the accuracy of the predictive models.

In Fig. 3, we have illustrated the percentage reduction in the number of features selected by each filtering method. The results show that the proposed filtering method

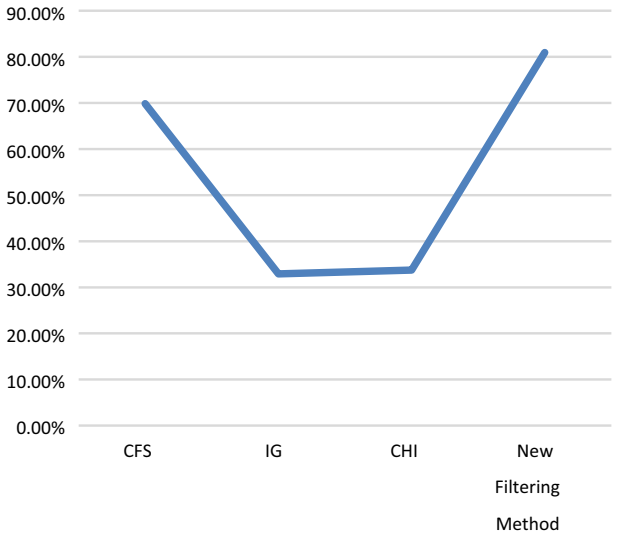

Fig. 3 Reduction in the number of features (given in percentage) derived on average for each filtering method over all datasets

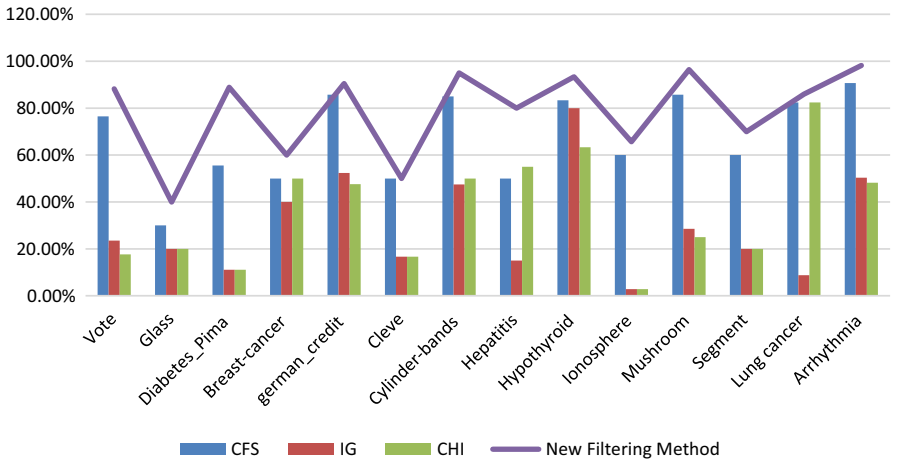

Fig. 4 Reduction in % of the number of features derived per dataset for the considered filtering method

is superior in reducing the dimensionality of the datasets by selecting, on average, fewer number of variables compared to the rest of the filtering methods.

To validate this finding, we have listed this reduction in percentage and per dataset as shown in Fig. 4.

Tables 3, 4 and 5 show the predictive models' performance based on the classification algorithms PART, eDRI, and C4.5 upon applying the filtering methods on the datasets. The impact of filtering methods on the predictive performance in the three data mining algorithms results is obvious. For instance, when the CFS filtering method

Table 3 Error rate (%) of the predictive models of PART after applying filtering methods

Dataset	All variables	IG	Chi	CFS	Our method
Vote	4.59	4.63	4.63	3.9	4.36
Glass	39.25	30.37	33.17	32.71	30.37
Diabetes_Pima	26.56	24.73	24.73	27.13	26.95
Breast-cancer	30.41	30.41	30.41	30.41	25.17
german_credit	30.7	26.9	29.3	28.5	30
Cleve	22.11	18.48	18.48	20.13	20.13
Cylinder-bands	41.48	40.74	41.48	41.11	42.22
Hepatitis	15.48	20.64	15.48	22.58	20
Hypothyroid	0.58	0.6	0.76	6.4	6.4
Ionosphere	8.26	8.26	8.26	10.82	11.11
Mushroom	0	0	0	0.98	1.47
Segment	3.76	3.54	3.54	7.01	7.01
Lung-cancer	43.75	25.00	28.12	34.37	59.38
Arrhythmia	35.61	42.69	39.82	41.15	43.80

Table 4 Error rate (%) of the predictive models of eDRI after applying the filtering methods

Dataset	All variables	IG	Chi	CFS	Our method
Vote	6.43	8.27	7.35	17.5	4.36
Glass	47.66	45.79	46.26	28.97	49.06
Diabetes_Pima	27.08	27.08	27.08	22.87	25.65
Breast-cancer	32.51	32.86	29.04	25.17	30.76
german_credit	28.8	27.9	30.9	28.6	32.04
Cleve	23.19	19.56	19.56	36.9	36.5
Cylinder-bands	8.51	25.18	40.74	7.4	7.4
Hepatitis	13.54	19.35	19.35	14.09	25.8
Hypothyroid	7.1	6.7	6.62	7.05	7.05
Ionosphere	7.12	13.67	33	5.41	6.55
Mushroom	0	0.29	0.29	0.98	1.47
Segment	4.19	12.42	12.42	3.5	8.44
Lung-cancer	50.00	37.50	43.75	46.87	78.13
Arrhythmia	69.91	69.91	74.55	84.51	34.73

was applied, the eight classifiers derived using eDRI algorithm have higher predictive power than those when no filtering was applied. The proposed filtering method employed during preprocessing shows highly competitive classifiers regardless of the data mining algorithm used. The win-lost-tie records of our proposed method when compared to IG, CHI and CFS for predictive models extracted by eDRI are 6-8-0, 5-9-0, and 2-10-2 respectively. When PART and C4.5 algorithms are used to generate the

Table 5 Error rate (%) of the predictive models of C4.5 after applying the filtering methods

Dataset	All variables	IG	Chi	CFS	Our method
Vote	3.67	3.67	3.67	3.9	4.36
Glass	42.05	32.17	31.3	31.77	31.3
Diabetes_Pima	26.17	26.17	26.17	25.13	25.52
Breast-cancer	24.47	24.82	26.92	26.92	26.92
german_credit	27.2	27.6	28.5	29.5	30
Cleve	24.42	24.42	24.42	22.11	22.11
Cylinder-bands	42.22	42.22	42.22	42.22	42.11
Hepatitis	16.12	20.64	17.41	20.64	23.87
Hypothyroid	0.42	0.58	0.66	6.68	6.44
Ionosphere	8.54	8.54	8.54	10.82	10.82
Mushroom	0	0	0	0.98	1.47
Segment	3.07	3.03	3.03	7.31	8.24
Lung-cancer	50.00	21.87	31.25	34.37	56.25
Arrhythmia	36.06	45.35	35.39	43.14	41.87

predictive models the win-lost-tie records of our method against IG, CHI and CFS are ("PART": 4-9-1, 4-10-0, 5-8-3) and ("C4.5": 5-9-0, 4-8-2, 4-6-4) respectively. These results show that the proposed method was able to exert a solid impact on the predictive models' performance and was indeed competitive to CHI, IG and CFS fileting methods.

Our filtering method was able to discard further insignificant variables when compared with other filtering methods without substantially hindering the classification accuracy of the predictive models. Hence a new competitive advantage has been created by the proposed filtering method that is achieved by balancing the number of required variables and the predictive performance. In other words, the proposed filtering method substantially cuts down the dimensionality of datasets in an exchange in some cases, with a slight decrease in accuracy. This is obvious from Fig. 5, which depicts the average predictive performance based on all datasets after applying the filtering methods. The graph illustrates that preprocessing not only reduces the feature space but is also highly accurate when compared to the results derived from the complete features set.

The average difference between the proposed method and (IG, CHI, CFS) with respect to prediction accuracy is -1.46, -0.41, and 0.68% respectively. These figures show that despite the substantial reduction in the number of variables the predictive models' performance was still good. The slight decrease in accuracy is minimal and has a limited impact when considering that there was more than a 50% further reduction in the datasets dimensionality. This indeed efficiently makes use of computing resources and, as we will see below, creates a concise set of rules.

We investigated the content of the predictive models produced by eDRI, PART and C4.5 algorithms based on the filtering methods. Our goal is to determine if the proposed method yields fewer rules without affecting the predictive performance.

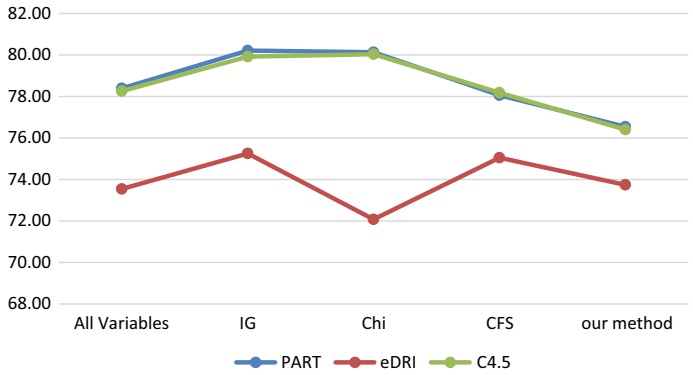

Fig. 5 The average predictive models accuracy of all filtering methods

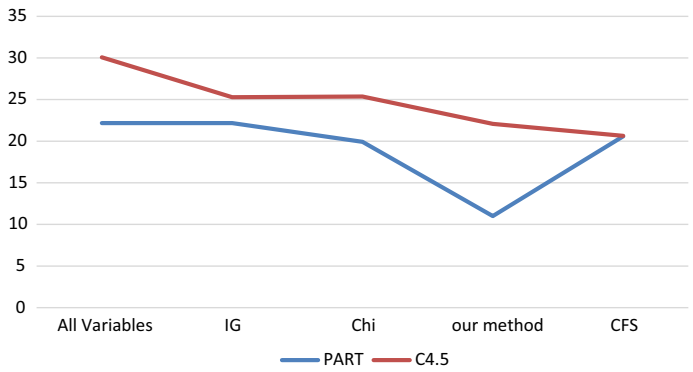

Fig. 6 The average number of rules produced by the predictive models of PART and C4.5 upon applying the filtering methods

Figure 6 displays the average number of rules derived by C4.5 and PART algorithms after applying the filtering methods. We have omitted eDRI results since this algorithm produced a large numbers of rules. Figure 6 shows that both C4.5 and PART produce a fewer number of rules when the features are filtered through our method than any of the benchmark filters. It is an expected outcome as our method filters out more variables than CHI, CFS and IG as was previously shown. The reduction in the number of rules produced by the classifiers allows for a better understanding of the underlying model used to make predictions.

Tables 6 and 7 give a more detailed view about the number of rules extracted by C4.5 and PART from each dataset upon applying the filters. In the case of the majority of the datasets considered, the classifiers derived by PART and C4.5 contain fewer rules with the exception of the "Segment" dataset. This dataset contains 20 continuous variables including the target class. The dataset is balanced with respect to class labels and there are seven class values. We conclude, based on the results shown in Tables 6 and 7, that number of rules in the predictive models decreased in case of each filtering method.

Table 6 Number of rules derived by the predictive models of PART after applying filtering methods

Dataset	All variables	IG	Chi	Our method	CFS
Vote	6	6	6	2	4
Glass	33	19	20	18	17
Diabetes_Pima	35	13	13	3	5
Breast-cancer	15	18	15	7	15
german_credit	69	72	79	1	17
Cleve	19	17	17	8	8
Cylinder-bands	17	50	17	1	51
Hepatitis	8	12	8	6	8
Hypothyroid	11	6	11	7	11
Ionosphere	10	10	10	12	12
Mushroom	13	13	13	7	14
Segment	29	30	30	70	85
Lung-cancer	7	4	3	5	5
Arrhythmia	99	15	35	41	5

Table 7 Number of rules derived by predictive models of C4.5 after applying the filtering methods

Dataset	All variables	IG	Chi	Our method	CFS
Vote	6	6	6	2	4
Glass	73	26	28	23	29
Diabetes_Pima	64	20	20	3	15
Breast-cancer	4	4	17	17	17
german_credit	90	80	54	1	15
Cleve	26	26	26	22	22
Cylinder-bands	1	1	1	1	1
Hepatitis	11	5	5	1	2
Hypothyroid	15	9	9	19	19
Ionosphere	18	18	18	37	37
Mushroom	25	30	30	9	17
Segment	39	39	39	70	39
Lung-cancer	16	7	13	4	16
Arrhythmia	33	83	89	100	56

5 Conclusions

Feature selection is an important part of any classification problem. Choosing the right features gives us a better insight into the problem and thus builds more robust classifiers. Minimizing the number of features also reduces the computational load of the classifier. In this paper, we propose a new two-step fileting method that uses the

combined scores of IG and CHI to refine the CFS subset. First, IG and CHI scores are normalized and combined into a single V-score. Then the new V-scores are used to filter out the low impact features from the CFS subset. The threshold for eliminating features is set at 50% of the highest V-score in the set. Thus, we were able to decrease the number of features used in the classification model with little effect on the accuracy of the resulting classifier.

In order to gauge the performance of our method we compared the accuracy of our feature selection method with that of IG, Chi and CSF methods using rule based predictive models. To this end, we ran each method on different datasets that belong to various application domains to obtain a robust comparison. One can see from the analysis in Sect. 4 that our method performed as well as other methods and often greatly improved on the existing methods with respect to data dimensionality reduction and predictive classifiers performance. Specifically, as shown in Table 2, our selection method generates a set of variables that is much smaller than IG, Chi squared and CSF subsets. It is in fact the greatest accomplishment of our method. We further showed that even with the reduced size of feature subset we are able to maintain the accuracy of the classifier (see Table 3). Lastly, the proposed filtering method when used with rule-based predictive models guarantees fewer numbers of rules without negatively impacting the predictive power of the resulting classifiers. Based on the above analysis, we believe that our proposed method for feature selection provides a technique that greatly reduces the number of features in the optimal subset while maintaining a robust level of accuracy and more controllable classifiers with fewer rules. Our method is computationally simple and practical.

In the future, we plan to test our fileting method in the context of unstructured datasets related to text categorization.

References

1. Abbasi A, France S, Zhang Z, Chen H (2011) Selecting attributes for sentiment classification using feature relation networks. IEEE Trans Knowl Data Eng 23(3):447–462
2. Abdelhamid N, Thabtah F, Ayesh A (2014) Phishing detection based associative classification data mining. Expert Syst Appl J 41:5948–5959
3. Al-Thubaity A, Abanumay N, Al-Jerayyed S, Alrukban A, Mannaa Z (2013) The effect of combining different feature selection methods on Arabic text classification. In: 2013 14th ACIS international conference software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD). IEEE, pp 211–216
4. Azhagusundari B, Thanamani AS (2013) Feature selection based on information gain. IJITEE 2(2):2278–3075
5. Dangi A (2013) Financial portfolio optimization: computationally guided agents to investigate, analyse and invest!? arXiv:1301.4194
6. Fahad A, Tari Z, Khalil I, Habib I, Alnuweiri H (2013) Toward an efficient and scalable feature selection approach for internet traffic classification. Comput Netw 57(9):2040–2057
7. Frank E, Witten I (1998) Generating accurate rule sets without global optimisation. In: Proceedings of the fifteenth international conference on machine learning, pp 144–151
8. Haddi E, Liu X, Shi Y (2013) The role of text pre-processing in sentiment analysis. Procedia Comput Sci 17:26–32
9. Hall M (1999) Correlation-based feature selection for machine learning. Waikaito University, thesis
10. Hoque N, Bhattacharyya DK, Kalita JK (2014) MIFS-ND: a mutual information-based feature selection method. Expert Syst Appl 41(14):6371–6385

11. Kunasekaran KK, Sugumaran R (2016) Exploratory analysis of feature selection techniques in medical image processing. In: Proceedings of the international conference on information engineering, management and security 2016 (ICIEMS 2016), pp 33–37
12. Lichman M (2013) UCI machine learning repository. University of California, Irvine. http://archive.ics.uci.edu/ml
13. Liu H, Setiono R (1995) Chi2: feature selection and discretization of numeric attribute. In: Proceedings of the seventh IEEE international conference on tools with artificial intelligence, pp 388–391
14. Mohammad R, Thabtah F, McCluskey L (2015) Tutorial and critical analysis of phishing websites methods. Comput Sci Rev J 17:1–24
15. Oreski S, Oreski G (2014) Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert Syst Appl 41(4):2052–2064
16. Parlar T, Özel SA (2016) A new feature selection method for sentiment analysis of Turkish reviews. In: Proceedings of the 2016 international symposium on innovations in intelligent systems and applications (INISTA)
17. Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106
18. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, Burlington
19. Thabtah F, Qabajeh I, Chiclana F (2016) Constrained dynamic rule induction learning. Expert Syst Appl 63:74–85
20. Thabtah F, Gharaibeh O, Al-zubaidy R (2012) Arabic text mining for rule based classification. J Inf Knowl Manag 11(1):1–10
21. Thabtah F, Kamalov F (2016) Phishing detection: a case analysis on classifiers with rules using machine learning. J Inf Knowl Manag (to appear)
22. Thabtah F, Abdelhamid F (2016) Deriving correlated sets of website features for phishing detection: a computational intelligence approach. J Inf Knowl Manag 15:1650042
23. Tsai CF, Hsiao YC (2010) Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches. Decis Support Syst 50(1):258–269
24. Uncu Ö, Türkşen IB (2007) A novel feature selection approach: combining feature wrappers and filters. Inf Sci 177(2):449–466
25. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington
26. Yang Y, Rogati M (2002) High-performing feature selection for text classification. In: Proceedings of the eleventh international conference on Information and knowledge management, pp. 659–661
27. Yousef M, Saçar MD, Khalifa W, Allmer J (2016) Feature selection has a large impact on one-class classification accuracy for MicroRNAs in plants. Adv Bioinform. doi:10.1155/2016/5670851
28. Zhou H, Guo J, Wang Y, Zhao M (2016) A feature selection approach based on interclass and intraclass relative contributions of terms. Comput Intell Neurosci. doi:10.1155/2016/1715780