Applied Soft Computing Journal 101 (2021) 107041

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

Extracting salient information from discarded features via attribute selection and pruning

Antonio J. Tallón-Ballesteros [a,*], Sung-Bae Cho [b]

[a] Department of Languages and Computer Systems, University of Seville, Seville, Spain
[b] Department of Computer Science, Graduate School of Artificial Intelligence, Yonsei University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 2 April 2020
Received in revised form 13 December 2020
Accepted 16 December 2020
Available online 18 December 2020

Keywords:
Classification
Feature selection
Data mining
Data preparation
Discarded attributes
Pruning

ABSTRACT

This paper introduces a novel two-step proposal for attribute subset selection. The first step is able to extract handy information from non-selected features after an initial feature subset selection to be added to the preliminary subset selected; for this step the feature subset selector relies on the Correlation-based Feature Selection (CFS). The second step tries to prune the solution given that in the first step some extra attributes out of initially non-selected features have been added to the relevant features; the ultimate goal of this pruning is to optimise the attribute subset prior to conduct a classification task. In accordance with the experimental results, the simplification of solutions is effective since an improvement takes place in most of the cases where the new method does determine some attributes to be removed. Some performance measures are reported within a test bed composed of a good number of binary and multi-class classification problems. Comparisons with other attribute selection procedures such as CFS, Fast Correlation-based Feature Selection and Gain Ratio revealed that the new approach is very competitive and may be taken into account by the data preparation community.

1. Introduction

Data Engineering (DE) is generally considered to be a central issue in the development of data mining applications [1]. The goal of DE is to put quality data on the hands of users [2]. Characteristics of problems play an important role in knowledge and data engineering applications since their solutions are clearly dependent. Problems can be classified according to three factors such as (a) the completeness of knowledge or data in the environment, (b) the accuracy of knowledge or data available in the environment and (c) the knowledge about the objective and/or requirements of the problem at hand [3]. The core procedures inside Machine Learning are categorised into Supervised Learning, Unsupervised Learning and Reinforcement Learning methods. The most widespread paradigm is Supervised Machine Learning (SML) [4]. This contribution focuses on SML where the objects are featured by the values for the properties and a label value is present in every sample [5]. Data Mining (DM) comprises three parts [6]: (a) Data Preparation (DP), (b) Data Surveying (DS) and (c) Data Modelling. Fig. 1 depicts the DM activities. DP also called Preparing the Data for modelling prepares data and also prepares the miner so that when using prepared data, the miner produces better models, faster; it cleans the records and also applies any kind of reduction at level of instances, features or classes; the results of this phase are training and testing sets. DS also called Surveying the data goals to look at the general structure and reports whether or not there is a useful amount of information enfolded in the data set about various areas; it tries to study the problem or even to detect some initial noise or inconsistencies on the data; it is conducted by a domain expert. Data Modelling involves a fundamental matter where the miner applies any modelling tool to produce a predictive model.

DP consumes approximately 60% of the total DE effort and is therefore a hot research topic [6]. It includes the tasks of data selection, data reorganisation, data exploration, data cleaning, and data transformation [7]. This paper pays special attention to Feature Selection (FS) which falls in the category of data selection. The pre-requirement to apply attribute selection is empty in the sense that having a few features in a problem is enough to carry out a data pre-processing task. In the past years, many publications with a very small input attribute space have been very useful to put on the table the necessity of feature selection such as [8–10] where some problems with three, four or eight initial properties (haberman, iris and pima data sets) are undergone attribute selection. Nowadays the number of features is on the rise due to use of many sensors and the very low cost of information storage. Another matter is how effective the solution to conduct a further learning procedure can be. As stated by very

[*] Corresponding author.
E-mail address: atallon@us.es (A.J. Tallón-Ballesteros).

https://doi.org/10.1016/j.asoc.2020.107041
1568-4946/© 2020 Elsevier B.V. All rights reserved.

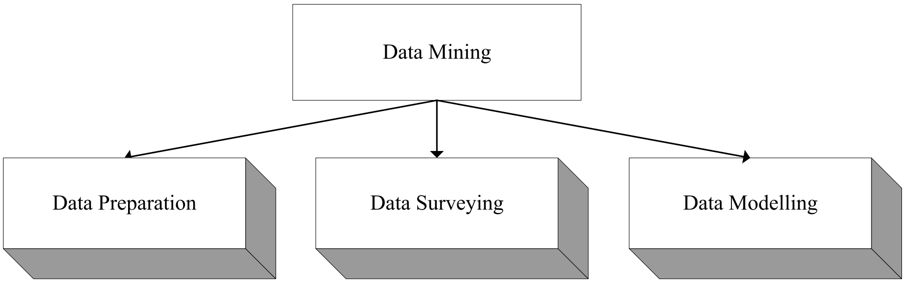


Fig. 1. Data Mining steps.

recent studies, a couple of features or even a single one may not be enough to foresee effectively on new, unseen data [11] with general purpose classifiers. Another issue is the case of single variable classifiers where the type of purpose is different and only one variable is necessary [12]. Feature selection can also be conducted to represent only important information about a particular situation in a 2D or 3D plot. In addition, other times the solution with feature selection may not be very accurate since there are many fields where real data are collected concerning scientific applications and, at the beginning, the purpose of the data gathering is not clearly in mind and some values may not be stored due to measure failures or even the quality and smartness of data not being the main goal [13].

This paper aims at optimising the attribute subset obtained with an advanced feature selection procedure [14]. The way to optimise is based on characteristics space pruning of the baseline solution. To do that, an attribute subset approach operating with a learning algorithm is run as an additional preparation phase on the solution initially reached; the important matter is that the target classifier is different from the inner learning method applied to optimise the solution. Moreover, the family of prediction models belongs to another approach to conduct a fair and unbiased study. As a result, the desirable end is to get more accurate models to predict with unknown samples. The scope of the proposed research falls into SML for classification tasks. The research question is motivated by the fact that a single Feature Subset Selector (FSS) may not be able to capture properly all the relationships among different attributes; to overcome this shortcoming Feature Subset Selection is populated with some extra attributes which are those selected applying a FSS to the discarded features in the first Feature Selection. The augmented solution contains essential and important features; although both are going to operate cooperatively, it is very likely that some attributes could be dropped without degrading the classifier performance. The contribution of this paper lies in the incorporation of a semi-wrapper FSS to optimise the baseline solution which is built with the application of an advanced Attribute Subset Selection (ASS) based on the classical Correlation-based Feature Selection proposed two decades ago. The fundamental motivation for using a semi-wrapper is its speed and acceptable performance compared to a wrapper. The rest of this manuscript is organised as follows. Section 2 reviews some concepts about attribute selection. Section 3 describes in detail the proposed method. Section 4 depicts the test bed and the experimentation framework. Section 5 shows the results for the experiments conducted to confirm the usefulness of the proposal. Section 6 states the concluding remarks and introduces future research lines.

2. Attribute selection

One of the cornerstones of DP is generating a data set smaller than the original one. Attribute Selection (AS) is a way for selecting relevant data via discarding some attributes that may not be very promising in terms of predictability [15]. The benefits of the AS application are varied and probably an increase in the model simplicity and the predictive power are the two of utmost importance. There are two main perspectives to perform AS: filters and wrappers. The former only makes use of inner properties of the data to determine any relationship between the attribute values and the associated label. The latter is guided by an induction learning to help pick up the highest quality solutions.

From another point of view, any AS approach may be divided into attribute ranking (AR) or attribute subset selection (ASS) depending on whether the procedure output is an ordered list or a subset [16]. Clearly, the computation time to get a score for every attribute is much more reduced than testing and assessing different combinations which includes typically as the first step getting an initial list and then picking one element and adding some of the subsequent elements of the list following a search strategy. AR is a linear complexity problem whereas ASS is a combinatorial problem that requires higher computational costs that may be an exponential time complexity for exhaustive search or a quadratic one for heuristic search.

Additionally, the ASS can be integrated within a wrapper and the running time may be a critical burden. Going further, the solutions provided by wrapper are too specific and therefore their applicability narrower. In the context of ASS, there is a very wide literature covering different search methods and different metrics to assess the subsets of attributes. It should be mentioned the complete review by B. Xue et al. which was published a few years ago and included more than two hundred references and is a very up-to-date survey of the state-of-the-art for ASS with heuristic-based approaches [17]. AR only requires a measure to analyse the dispersion, correlation, mutual information (also called entropy), chi-squared statistics, symmetrical uncertainty, etc. of the data where the attributes are projected one at a time along with the class label. A novelty work combines simultaneously AS with discretisation techniques conducting experiments with some AR methods [18]. In accordance with a recent survey [19], current FS techniques primarily concentrate on obtaining relevant attributes.

ASS is a branch of AS with the highest number of publications within the data mining community probably due to its good performance. The most widespread method is Correlation-based feature Selection (CFS) [20] which incorporates a Best First search along with a correlation measure to assess the quality of the subsets. The usual is to discard non-selected attributes and to only include the selected attributes to train the supervised machine learning model. A new proposal introduced the application of an ASS method even to the non-selected attributes and in a subsequent step the selected attributes from the original training set and the selected ones from the non-selected initially are merged and are the input to the classification algorithm [14]. From a different perspective, FS may be tackled with static FS and streaming FS: in the former the original data do not change over the time and in the latter new features may be added to the original data [21].

Table 1
Test bed summary.

Data set	#Instances			#Attributes	#Labels	Attribute-to-label ratio
	Total	*Training*	*Testing*			
AID1608	1033	827	206	152	2	76.0
B. tissue	106	81	25	9	6	1.5
CTG	2126	1594	532	22	3	7.3
Led24	3200	200	3000	24	10	2.4
MADELON	2000	1500	500	500	2	250.0
magic	19020	14265	4755	10	2	5.0
SPECTF	267	80	187	44	2	22.0
STAD	100	75	25	14	3	4.7
Waveform	5000	3750	1250	40	3	13.3
Vowel	990	528	462	11	11	1.0
Average	3384.2	2290.0	1094.2	82.6	4.4	38.3
Range	100–19020	75–14265	25–4755	9–500	2–11	1-250

3. The proposed method

This paper presents a method to optimise the solution reached by an advanced attribute subset selection procedure by means of CFS; the reason of this choice is motivated by the fact that CFS is considered a very powerful method all over the data mining community and it counts with around two thousands and eight hundreds citations according to scholar google (http://scholar.google.com) at the moment of this paper submission. CFS considers the hypothesis that good attribute subsets contain attributes with a high correlation with the label, yet uncorrelated with each other [20]. The starting point is the initial solution which is submitted to an additional CFS with a different characteristic space; both attribute sets are merged and we call it baseline scenario.

The new approach carries out an additional attribute subset selection with Boruta algorithm [22]. Boruta applies Random Forest (RF) classifier to determine all-relevant feature subset from problems and is designed as a wrapper approach. There are some advantages to apply RF [23] as the engine of the wrapper such the number of parameters to be set up is very small [22] and the classification performance is very good. Boruta arises from the spirit of RF and hence adding randomness to the system several; basically, only variables whose importance is higher than that of the randomised variables are considered important. Furthermore, RF runs are performed several times computing the importance of all attributes [22]. The operation of Boruta [24] is as follows: (i) it evaluates the relevance of variable in the input data by comparing the importance measured by RF for original attributes with that obtained for the random features added in an artificial fashion; to do that, it trains RF on the problem extended with random contrast attributes, obtained by shuffling randomly values of original features between instances, ii) for every feature, it is tested if its importance is higher than the maximal importance achieved by any contrast attribute; in order to obtain statistically significant results the procedure is looped several times, with contrast variables generated independently for each iteration. Technically, Boruta is a wrapper method. Nonetheless, this contribution does not consider RF as the target classifier and therefore we can state that Boruta is applied as a semi-wrapper FS method [25]. Thus, the proposed method mixes an advanced attribute selection procedure based on a sophisticated application of CFS with Boruta as a semi-wrapper and thus the new approach falls into a new type of attribute selection procedure that we may call hybrid filter and semi-wrapper attribute subset selection.

Moreover, this paper considers as target classifiers some induction algorithms belonging to strategies that do not follow decision trees specially to analyse the convenience of the proposed method as a general attribute subset selection framework. The rationale to conduct Boruta only with an initial tentative good quality solution is to diminish the computational cost and at the same time to study if some attributes can be safely discarded because the initial solution is bigger than the solution with the application of CFS only once. Thus, Boruta is used as an optimiser. The advantages of using a classifier to assess the quality of a solution are the higher accuracy and also that a type of classification algorithm has been applied before carrying out the learning phase with unseen data.

Fig. 2 portrays the proposed method composed of two parts. The first phase called *advanced attribute selection* performs a feature subset selection via CFS in two sequential steps (double CFS), considering the original training data and the complementary reduced training data as inputs of the aforementioned steps, respectively. The second phase named *optimisation* takes the combination of the solutions achieved by each of the selections conduced in the first phase and tries to optimise the solution via Boruta algorithm which makes use of RF to assess the solution quality. With regard to the sketch, any classifier may be used for the subsequent learning process although this paper only evaluates the solution reached with learning algorithms which do not fall within the category of inductive decision trees. The performance of baseline and optimised scenarios are going to be compared and it may help to measure the behaviour of our proposal. Fig. 3 presents the pseudo-code of the proposal. In the first phase, two attribute subsets via CFS are obtained and conveniently combined since there is no overlapping; the most relevant of the non-selected attributes are retained for a further processing. In the second phase an optimisation of the solution obtained in the first step takes place via Boruta algorithm.

In the specialised literature there are a few attempts to address a two-stage attribute selection. A two-phase approach applying an embedded method based on Support Vector Machines Recursive Feature Elimination (SVM-RFE) is proposed in [26]. More recently, S. Jadhav et al. [27] presented a method conducting a Feature Ranking followed by a wrapper approach in credit rating context. In 2020, P. K. Theodorikis and D. C. Gkikas [28] describe a model based on a wrapper approach combined with decision trees although no empirical analysis is conducted.

4. Experimentation setting

Table 1 gives details about the test bed to assess the proposed method which is composed of ten binary and multiple class problems. The data sets are very interesting since they represent complex real-world or challenge problems; most of them are available at the international well-known repository from the University of California at Irvine [29], MADELON has been introduced in NIPS 2003 challenge [30] and STAD that stands for STomach Adenocarcinoma is a problem from Bioinformatics [31]. There are six multiple class data sets and the remaining are binary, throwing an average of 4.4 classes. The size in terms of instances varies from 100 to 19020. The attribute space size goes

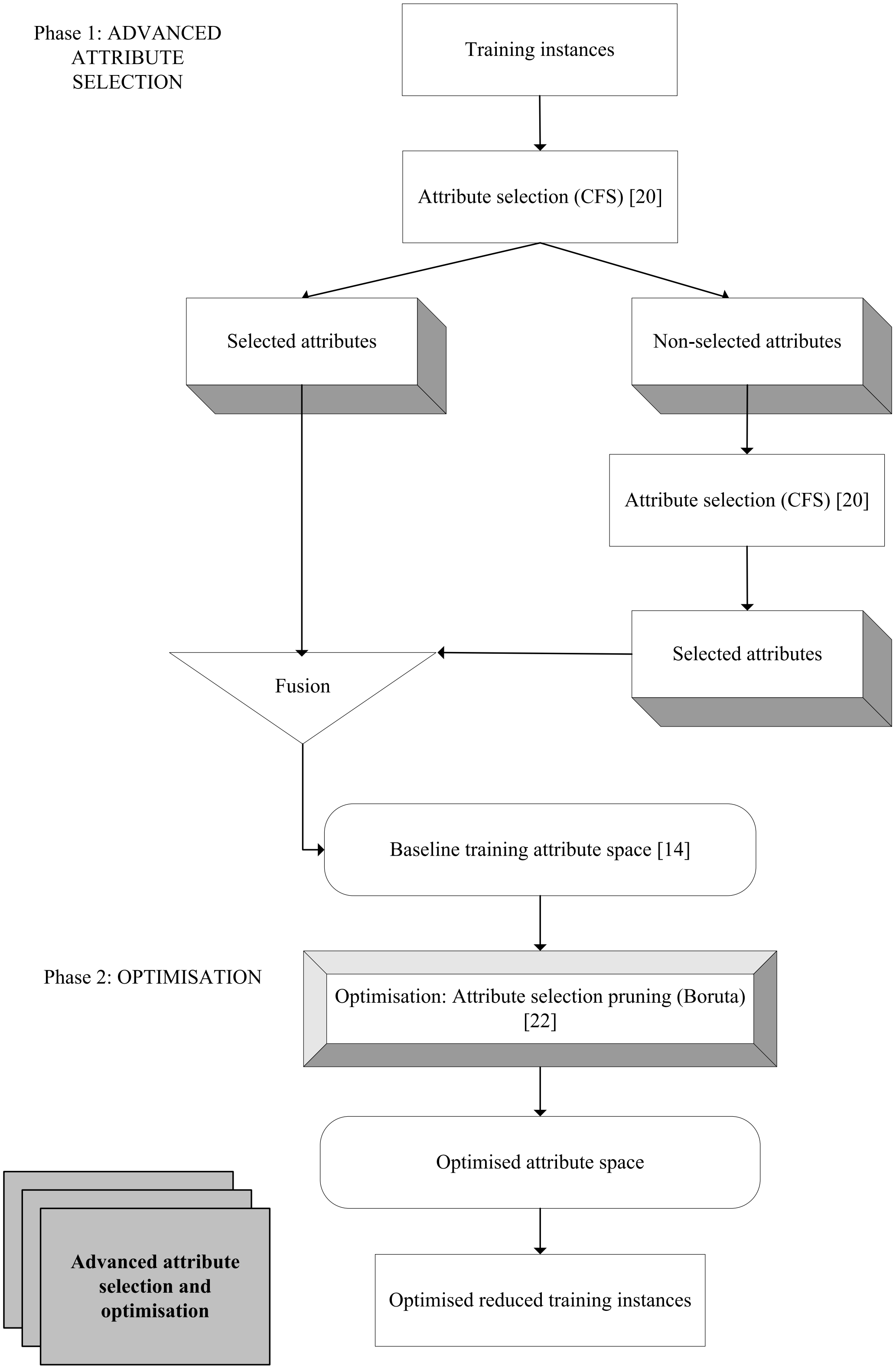
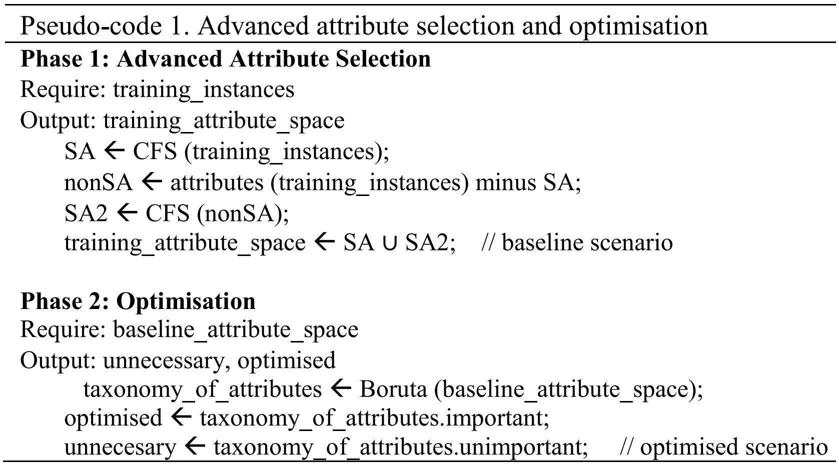

Phase 1: ADVANCED ATTRIBUTE SELECTION
Training instances
Attribute selection (CFS) [20]
Selected attributes
Non-selected attributes
Attribute selection (CFS) [20]
Selected attributes
Fusion
Baseline training attribute space [14]
Phase 2: OPTIMISATION
Optimisation: Attribute selection pruning (Boruta) [22]
Optimised attribute space
Advanced attribute selection and optimisation
Optimised reduced training instances

Fig. 2. The proposal framework: Advanced attribute selection and optimisation.

from 9 to 500. Last column presents the Attribute-to-label ratio to provide a measure between the attribute space size and the number of labels which may be interpreted as a measure of how much attributes are required from every label on average; this measure is reported throughout this contribution and enables us to get a measure about the simplification and eventual optimisation that the proposal has reached with an extra quantitative value. The experimental design follows a stratified holdout procedure [32] with three and one quarters for the training and testing set, respectively. There are some particular cases where the problems were pre-arranged and we keep on this distribution; some examples are Led24 [33], SPECTF [34] and vowel [35].



```
Pseudo-code 1. Advanced attribute selection and optimisation
Phase 1: Advanced Attribute Selection
Require: training_instances
Output: training_attribute_space
    SA ← CFS (training_instances);
    nonSA ← attributes (training_instances) minus SA;
    SA2 ← CFS (nonSA);
    training_attribute_space ← SA ∪ SA2;    // baseline scenario

Phase 2: Optimisation
Require: baseline_attribute_space
Output: unnecessary, optimised
      taxonomy_of_attributes ← Boruta (baseline_attribute_space);
    optimised ← taxonomy_of_attributes.important;
    unnecesary ← taxonomy_of_attributes.unimportant;    // optimised scenario
```

Fig. 3. Pseudo-code of the proposal *Advanced attribute selection and optimisation*.

Table 2
Attribute selection methods and classifiers.

Name	Type	Parameter	Value	Target classifier
Correlation based Feature Selection (CFS)	Filter	Attribute evaluation measure Type of search Consecutive expanded nodes without improving Direction of search	Correlation Best first 5 Forward	1NN, PART and NB
Boruta	Semi-wrapper	Classifier to assess optimisation	Random forest	1NN, PART and NB

Table 2 enumerates the approaches used within the proposed method in the phases 1 and 2 along with a brief characterisation and also including their setting and last column represents the target classifiers. The results obtained after the double CFS are considered as the baseline ones and those raised after the application of both stages, named double CFS and Boruta, are the optimised results. Concerning the technology for developing the proposed framework we have used some visual tools [36] such as Weka (Waikato Environment for Knowledge Analysis) [37] and RStudio [38] which are written in Java and R programming languages.

For CFS and Boruta, the implementations of Weka and R package Boruta [39] within RStudio have been used, respectively. The target classifiers are *k*-Nearest Neighbour (since in this contribution k=1, it is referred as to 1NN) [40] which is based on the nearest neighbour rule, PART which is a method to obtain rules from PARtial decision Trees [41] and Naïve Bayes (NB) [42], which is supported by the Bayes Theorem, have been run under Weka framework. We have set up the AS and learning algorithms with the default parameters that are those suggested by the own authors of the algorithms.

Table 3 represents the number of attributes in the initial situation, after applying the first phase of the proposed method (baseline scenario) and after being conducted fully the new proposal. It is worth mentioning that in some cases, the proposal is not able to reduce the solution reached with the advanced attribute selection which may mean the high quality of it. Last column reports the attribute-to-label ratio of the baseline and optimised situations; concerning the numerical values, we can see that, on average, around 3 or 4 attributes for every label are required to conduct the learning process. In addition, no more than eight features per class are required for binary problems and no more than six attributes, per label too, to cope with multiple class data sets. Furthermore, the number of unnecessary features from the baseline solution ranges from 1 to 5 in the cases where is possible to optimise the solution. In general terms, the theoretical lower bound for the selected attributes is one; from an empirical point of view, the practical lower bound to get reliable results is two [10]. Table 4 illustrates a reduction analysis on data sets where the proposed method is able to optimise the solution with the baseline approach. The reduction percentage with the baseline strategy is very remarkable with values from 18.2 to 96.8 with an average of 64.1. Moreover, the optimisation can reduce even a bit more the baseline solutions with improvements that fluctuate from 0.8% to 14.3% raising an average of 5.9. The important news are the difference between 100% and the reported value on column labelled *Optimised over initial* which is the global optimisation measure, raising numbers from 2.0% (100.0%–98.0% for AID1608) to 81.3% (100.0%–22.7% for CTG) and an average of 29.9% (100.0%–70.1% for the whole test bed) which represents the percentage of needed features from the full data set. Consequently, only the problems that have been shown in Table 4 are going to be dealt with hereinafter.

As performance measures the accuracy and Cohen's kappa have been chosen to evaluate the models obtained by the classification models, once the classifiers are trained. Cohen's kappa measure is of special interest for multi-class problems [43]. Accuracy is a very well-known measure to score the number of hits divided by the number of samples to be assessed. Cohen's kappa is a convenient alternative measure since it compensates for random hits [44]. Given that all the considered attribute selection methods and classifiers are non-stochastic algorithms, the number of runs is one and the test result is reported on the forthcoming tables.

5. Results

This section illustrates the results for the baseline and optimised scenarios with three classifiers, which are described in Section 4, and two performance measures with baseline and optimised reduced testing instances. Lastly, we present the results with other additional attribute selection methods together with a global plot of them for the two assessment metrics reported in this paper.

Table 3
Selected Attributes in the baseline scenario and with the Proposed Approach along with the attribute-to-label ratio.

Data set	#Attributes				#Labels	Attribute-to-label ratio		
	Initial	*Baseline*	*Optimised*	*Unnecessary*		*Initial*	*Baseline*	*Optimised*
AID1608	152	6	3	3	2	76.0	3.0	1.5
B. tissue	9	8	8	–	6	1.5	1.3	1.3
CTG	22	18	17	1	3	7.3	6.0	5.7
Led24	24	7	7	–	10	2.4	0.7	0.7
MADELON	500	16	12	4	2	250.0	8.0	6.0
magic	10	6	6	–	2	5.0	3.0	3.0
SPECTF	44	20	15	5	2	22.0	10.0	7.5
STAD	14	5	3	2	3	4.7	1.7	1.0
Waveform	40	18	17	1	3	13.3	6.0	5.7
Vowel	11	10	10	–	11	1.0	0.9	0.9
Average	82.6	11.4	9.8	2.7	4.4	38.3	4.1	3.3
Range	9–500	5–20	3–17	1–5	2–11	1–250	0.7–10.0	0.7-7.5

Table 4
Reduction analysis on problems where the Proposed Approach is able to optimise the baseline solution.

Data set	#Attributes					Reduction percentage (%)		
	Initial	*Baseline*	*Optimised*	*Unnecessary from optimised*	*Unnecessary percentage (%)*	*Baseline over initial*	*Optimised over initial*	*Improvement*
Optimised over baseline								
AID1608	152	6	3	3	50.0	96.1	98.0	2.0
CTG	22	18	17	1	5.6	18.2	22.7	4.5
MADELON	500	16	12	4	25.0	96.8	97.6	0.8
SPECTF	44	20	15	5	25.0	54.5	65.9	11.4
STAD	14	5	3	2	40.0	64.3	78.6	14.3
Waveform	40	18	17	1	5.6	55.0	57.5	2.5
Average	128.7	13.8	11.2	2.7	25.2	64.1	70.1	5.9
Range	14–500	5–20	3–17	1–5	5.6–50.0	18.2–96.8	22.7-98.0%	0.8-14.3%

Table 5
1NN classifier: test results.

Data set	Accuracy			Cohen's kappa		
	Baseline	*Optimised*	*Improvement*	*Baseline*	*Optimised*	*Improvement*
AID1608	93.20	93.69	*0.49*	−0.0091	0.0000	*0.0091*
CTG	80.26	80.26	0.00	0.4802	0.4802	0.0000
MADELON	71.00	83.60	*12.60*	0.4200	0.6720	*0.2520*
SPECTF	62.57	63.64	*1.07*	0.1211	0.1278	*0.0067*
STAD	56.00	72.00	*16.00*	0.2403	0.4957	*0.2553*
Waveform	74.00	75.20	*1.20*	0.6079	0.6259	*0.0180*
Average	72.84	78.06	5.23	0.3101	0.4003	0.0902
W/T/L			5/1/0			5/1/0

Table 6
PART classifier: test results.

Data set	Accuracy			Cohen's kappa		
	Baseline	*Optimised*	*Improvement*	*Baseline*	*Optimised*	*Improvement*
AID1608	93.69	93.69	0.00	0.0000	0.0000	0.0000
CTG	81.95	81.95	0.00	0.5212	0.5212	0.0000
MADELON	62.60	64.40	*1.80*	0.2520	0.2880	*0.0360*
SPECTF	72.19	72.19	0.00	0.1942	0.1942	0.0000
STAD	36.00	60.00	*24.00*	−0.0499	0.1639	*0.2137*
Waveform	76.80	78.56	*1.76*	0.6514	0.6773	*0.0259*
Average	70.54	75.13	4.59	0.2615	0.3074	0.0459
W/T/L			3/3/0			3/3/0

5.1. Results for the baseline and optimised scenarios

The results obtained for the baseline (only phase 1 is done) and optimised (this proposal, where phases 1 and 2 are conducted) scenarios for different classifiers are presented. The improvement values have been highlighted in italics. Table 5 presents the results obtained with 1NN classifier. The accuracy has been enhanced in five out of six problems with an average improvement of 5.23. Cohen's kappa is increased on average 0.0902 which means a 9%. A tie happens in CTG although the optimisation is interesting given that an attribute that has been featured unnecessary according to Boruta is also unnecessary in accordance with the results and the model at the end is simpler and faster to train. 1NN classifier takes important advantages with the proposal and may improve the typical storage problems that suffer classifiers based on nearest neighbours. MADELON has improved outstandingly and has passed from an accuracy of 71.00% to 83.60%.

Table 6 reports the performance of the proposed method with PART. The progress now takes place in three out of six problems and the improvements are a bit lower than with 1NN. The ties

Table 7
NB classifier: test results.

Data set	Accuracy			Cohen's kappa		
	Baseline	*Optimised*	*Improvement*	*Baseline*	*Optimised*	*Improvement*
AID1608	93.20	93.69	*0.49*	−0.0091	0.0000	*0.0091*
CTG	82.33	83.08	*0.75*	0.5171	0.5418	*0.0247*
MADELON	57.40	57.40	0.00	0.1480	0.1480	0.0000
SPECTF	71.12	70.59	−0.53	0.2238	0.2189	−0.0049
STAD	64.00	72.00	*8.00*	0.3680	0.4957	*0.1277*
Waveform	78.48	78.56	*0.08*	0.6788	0.6800	*0.0012*
Average	74.42	75.89	1.46	0.3211	0.3474	0.0263
W/T/L			4/1/1			4/1/1

Table 8
A comparison of global test accuracy and Cohen's kappa results for several attribute selection methods with classifiers 1NN, PART and NB.

Data set	Attribute selection method	#Attributes	**Accuracy**			**Cohen's kappa**		
			Classifier			*Classifier*		
			1NN	*PART*	*NB*	*1NN*	*PART*	*NB*
AID1608	FCBF	4	93.20	**93.69**	**93.69**	−0.0091	**0.0000**	**0.0000**
	CFS	5	93.20	**93.69**	**93.69**	-0.0091	**0.0000**	**0.0000**
	GainRatio	3	93.20	**93.69**	**93.69**	-0.0091	**0.0000**	**0.0000**
	Optimised	3	**93.69**	93.69	93.69	**0.0000**	**0.0000**	**0.0000**
CTG	FCBF	8	**81.20**	77.26	82.52	**0.4931**	0.3790	0.5283
	CFS	7	80.45	81.20	82.89	0.4773	0.4598	0.5413
	GainRatio	17	79.70	**82.14**	82.89	0.4895	**0.5229**	0.5340
	Optimised	17	80.26	81.95	**83.08**	0.4802	0.5212	**0.5418**
MADELON	FCBF	7	56.00	60.40	**57.80**	0.1200	0.2080	**0.1560**
	CFS	12	65.80	60.80	56.20	0.3160	0.2160	0.1240
	GainRatio	12	69.60	**65.40**	56.80	0.3920	**0.3080**	0.1360
	Optimised	12	**83.60**	64.40	57.40	**0.6720**	0.2880	0.1480
SPECTF	FCBF	6	59.36	64.71	68.45	0.0844	0.1349	0.1621
	CFS	12	57.75	70.05	68.98	0.0763	0.1543	0.1860
	GainRatio	15	59.36	71.12	67.38	0.0844	0.1845	0.2091
	Optimised	15	**63.64**	**72.19**	**70.59**	**0.1278**	**0.1942**	**0.2189**
STAD	FCBF	4	64.00	52.00	60.00	0.3284	0.1501	0.2877
	CFS	4	64.00	52.00	60.00	0.3284	0.1501	0.2877
	GainRatio	3	**72.00**	**60.00**	**72.00**	**0.4957**	**0.1639**	**0.4957**
	Optimised	3	**72.00**	**60.00**	**72.00**	**0.4957**	**0.1639**	**0.4957**
Waveform	FCBF	5	69.12	74.00	78.00	0.5341	0.6089	0.6688
	CFS	14	**75.36**	77.04	**80.00**	**0.6288**	0.6556	**0.7011**
	GainRatio	17	74.48	78.40	77.60	0.6152	0.6747	0.6656
	Optimised	17	75.20	**78.56**	78.56	0.6259	**0.6773**	0.6800
Average	FCBF	5.67	70.48	70.34	73.41	0.2585	0.2468	0.3005
	CFS	9.00	72.76	72.46	73.63	0.3030	0.2726	0.3067
	GainRatio	11.17	74.72	**75.13**	75.06	0.3446	**0.3090**	0.3401
	Optimised	11.17	**78.06**	**75.13**	**75.89**	**0.4003**	0.3074	**0.3474**
Global individual wins	FCBF		1	1 (*)	2 (1*)	1	1 (*)	2 (1*)
	CFS		1	1 (*)	2 (1*)	1	1 (*)	2 (1*)
	GainRatio		1 (*)	**4 (2*)**	2 (2*)	1 (*)	**4 (2*)**	2 (2*)
	Optimised		**4 (1*)**	**4 (2*)**	**4 (2*)**	**4 (1*)**	**4 (2*)**	**4 (2*)**
W/L for Optimised versus	FCBF		5/1	5/0	4/1	5/1	5/0	4/1
	CFS		4/2	5/0	4/1	5/1	5/0	4/1
	GainRatio		5/0	2/2	4/0	4/1	2/2	4/1

* Means ex-aequo.

go for AID1608, CTG and SPECTF. The number of attributes in AID1608 has been divided by two which represents an important reduction in terms of complexity and the final attribute space is able to keep the results. In SPECTF a quarter of the feature space has been discarded which is very remarkable. CTG has been mentioned with the previous classifier though now the accuracy and Cohen's kappa are higher than with 1NN which may mean that could be an improvement margin with 1NN classifier probably with another optimiser.

Table 7 exhibits the results obtained with NB classifier. The number of wins is four out of six. There is one tie and one loss. The loss is the first time that has happened with the proposal and perhaps may indicate that the semi-wrapper approach based on RF is not a very suitable option for NB or even that the SPECTF problem requires a further analysis and any noise may be present in the data. The average is improved only 1.46% in accuracy and 0.0263 in Cohen's kappa which is lower than with PART even though now there is one extra win and one tie has been changed to loss. MADELON has resulted in a tie although it is positive since the reduction in the attribute space is around a quarter.

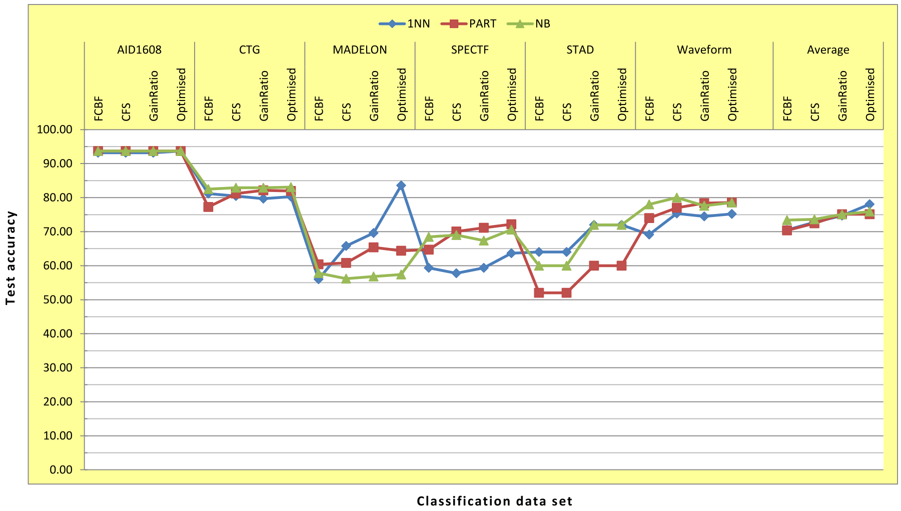


Fig. 4. Test accuracy plot with global results for our approach (*Optimised*) compared to three attribute selection procedures for three classifiers.

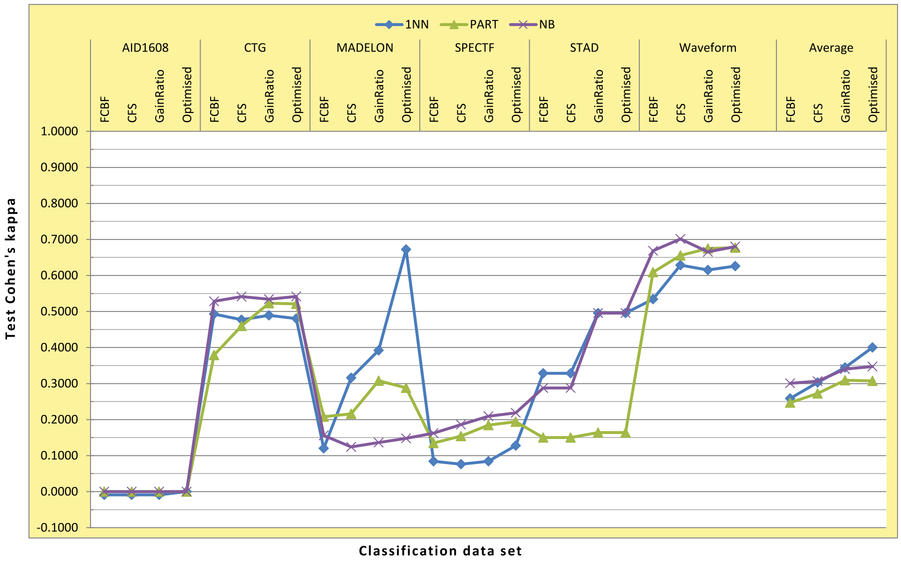


Fig. 5. Test Cohen's kappa plot with global results for our approach (*Optimised*) compared to three attribute selection procedures for three classifiers.

5.2. Results obtained with various attribute selection procedures

A comparison applying the current proposal (advanced attribute selection and optimisation; it is also referred as to *optimised* -scenario-), *Fast Correlation-based Feature Selection* (FCBF) [45] which is an information theory-based method, CFS which is a statistics-based method and *Gain Ratio* [46] as Feature Ranking based on Gain Ratio, has been reported. We have tested with the same classifiers of the previous subsection to get a fair comparison with the proposal.

Regarding the parameters, the algorithms have been run with the Weka default values because those are the recommended by the own authors of the algorithms, with the exception of Gain Ratio which has been setup to retained the same number of attributes as our proposal selected. Table 8 compares the performance results with FCBF, CFS, Gain Ratio and our proposed approach as well as the number of selected attributes. The best result appears in boldface for every pair (data set, classifier) as well as for the summary rows (entitled Average and Global individual wins) to indicate which the recommended attribute selection method is in connection with the empirical study; the last part of the aforementioned table does not highlight any element as it is a pairwise comparison where the proposal (named optimised inside the cell) is the control method. The analysis of the results from a qualitative point of view, bearing in mind the global individual wins and the ex-aequo results suggests that the new approach is very competitive since for the classifiers 1NN and NB the higher number of wins falls into the Optimised scenario and for PART the selector Gain Ratio is able to achieve very good results. As a follow-up comment on PART, it is worth mentioning that on average Gain Ratio and Optimised data preparation methods have the same result for accuracy with the significance of two decimal positions although with a precision of three decimal the tie goes in favour of the Optimised one; contrarily, for Cohen's kappa measure the best mean value is for Gain Ratio. The pairwise comparison in terms of Wins and Losses (W/L) excluding the ties, reveals that our proposal versus FCBF gets between 4 and 5 wins and hence 1 or 0 losses, respectively, for every classifier with all the measures; optimised versus CFS wins in every scenario although with scores such as 4/2, 4/1 or 5/0 for accuracy and 4/1, 5/1 or 5/0 for Cohen's kappa which means that the new approach is slightly more stable for Cohen's kappa measure. Our proposal versus Gain Ratio achieves important improvements for 1NN and NB classifiers and for PART a tie which balances equally both feature selection methods although it is important to remark that the new approach does not need to establish a cut-off value to select a concrete number of attributes and by its part Gain Ratio does need one and to do that the application of our contribution is required what makes dependent the Gain Ratio operation of the optimised feature selection procedure. The number of attributes selected by FCBF and CFS is the lowest although the performance is not very promising; it shows that there is a trade-off between the complexity and the performance.

Fig. 4 shows the test accuracy results for every data set as well as for the average in the four considered situations, id est, the FCBF, CFS, Gain Ratio and optimised feature selection methods for the three reference classifiers; generally speaking, the trend is increasing, from FCBF, to CFS, to Gain Ratio to optimise approach, which means that the individual performance for almost every data set fits the general averaged behaviour of the classifier for the whole test bed. It is obvious that for MADELON the 1NN classifier is a very appropriate choice and in particular the optimised data preparation method allows the classifier to take an important advantage; as it can be seen for SPECTF the classifiers PART and NB are preferable to 1NN one. For STAD, the PART classifier has a poor performance.

From Fig. 5 it is clear that the optimised approach yields a strong return for the Cohen's kappa and the exception of MADELON along with Gain Ratio, the proposal contributes to increase the classifier performance compared to the other three approaches. For MADELON and Waveform, the Cohen's kappa in the optimised scenario surpasses 0.6; especial attention deserves MADELON where the initial values are in the range [0.1–0.2].

The results have been plotted in Figs. 4 and 5, concerning the test accuracy and Cohen's kappa, respectively, to see clearly the global performance.

6. Conclusions

This paper introduced a proposal to optimise the attribute subset selection with a semi-wrapper feature selection based on advanced inductive decision trees such as Random Forest. The new approach has been validated on six multiple class and four binary classification problems. The average number of unnecessary attributes from the optimised scenario is 2.7 with minimum and maximum values of 1 and 5, respectively; it means a percentage of unnecessary features about 25.2%. The final reduction percentage over the initial context is, on average, 70.1% which is interpreted that only around a 30% of the initial inputs are necessary to have enough information to train the classifier with warranties in terms of predictive power.

1NN performance has been substantially enhanced with the proposed method and there is a clear advantage compared to the baseline approach in five out of six problems. PART classifier has overcome the results in three situations and the remaining ones are a tie which is also very remarkable since a reduction in the model complexity has been achieved without sacrificing performance. Lastly, NB classifier got four wins, one tie and one loss although the average enhancement is very reduced and represents around a 1.4% in accuracy and 2% in Cohen's kappa.

Our proposal compared to three feature selection approaches based on attribute subsets or attribute ranking is very competitive and may be seriously considered by the data mining community.

As the future work, we plan to conduct an empirical study on high-dimensional real-world Bioinformatics problems and big data samples especially those with millions of features. Another interesting future research topic could be to speed up a wrapper with a prior semi-wrapper; this may reduce the computational cost and hence exploring more widely the approximate baseline attribute space.

CRediT authorship contribution statement

Antonio J. Tallón-Ballesteros: Conceptualization, Software, Writing - review & editing. **Sung-Bae Cho:** Formal analysis, Software, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been partially supported by TIN2017-88209-C2-R project of the Spanish Inter-Ministerial Commission of Science and Technology (MICYT) and FEDER funds, and also by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2020-0-01361, Artificial Intelligence Graduate School Program (Yonsei University)).

References

[1] M.A. Hall, G. Holmes, Benchmarking attribute selection techniques for discrete class data mining, IEEE Trans. Knowl. Data Eng. 15 (6) (2003) 1437–1447.
[2] R.Y. Wang, H.B. Kon, S.E. Madnick, Data quality requirements analysis and modeling, in: Data Engineering, 1993. Proceedings. Ninth International Conference on, IEEE, 1993, pp. 670–677.
[3] C.V. Ramamoorthy, B.W. Wah, Knowledge and data engineering, IEEE Trans. Knowl. Data Eng. 1 (1) (1989) 9–16.
[4] M.I. Jordan, T.M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science 349 (6245) (2015) 255–260.
[5] S. Marsland, Machine Learning: An Algorithmic Perspective, CRC press, 2015.
[6] D. Pyle, Data Preparation for Data Mining, Vol. 1, Morgan Kaufmann, 1999.
[7] M. Kamel, Data preparation for data mining, in: Encyclopedia of Data Warehousing and Mining, second ed., IGI Global, 2009, pp. 538–543.
[8] Jr L.F. Guseman, B.P. Marion, LFSPMC: Linear feature selection program using the probability of misclassification, 1975.
[9] E. Cantú-Paz, Feature subset selection, class separability, and genetic algorithms, in: Genetic and Evolutionary Computation Conference, Springer, Berlin, Heidelberg, 2004, pp. 959–970.
[10] M. Prasad, A. Sowmya, I. Koch, Efficient feature selection based on independent component analysis, in: Intelligent Sensors, Sensor Networks and Information Processing Conference, 2004. Proceedings of the 2004, IEEE, 2004, pp. 427–432.
[11] A.J. Tallón-Ballesteros, J.C. Riquelme, Low dimensionality or same subsets as a result of feature selection: an in-depth roadmap, in: International Work-Conference on the Interplay Between Natural and Artificial Computation, Springer, Cham, 2017, pp. 531–539.
[12] H. Altınçay, Feature extraction using single variable classifiers for binary text classification, in: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, Berlin, Heidelberg, 2013, pp. 332–340.
[13] E. Cantú-Paz, S. Newsam, C. Kamath, Feature selection in scientific applications, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2004, pp. 788–793.
[14] A.J. Tallón-Ballesteros, L. Correia, B. Xue, Featuring the attributes in supervised machine learning, in: International Conference on Hybrid Artificial Intelligence Systems, Springer, Cham, 2018, pp. 350–362.
[15] S. Zhang, C. Zhang, Q. Yang, Data preparation for data mining, Appl. Artif. Intell. 17 (5–6) (2003) 375–381.
[16] H. Jia, et al., Data transformation and attribute subset selection: Do they help make differences in software failure prediction?, in: Software Maintenance, 2009. ICSM 2009. IEEE International Conference on, IEEE, 2009, pp. 519–522.
[17] B. Xue, et al., A survey on evolutionary computation approaches to feature selection, IEEE Trans. Evol. Comput. 20 (4) (2016) 606–626.
[18] M.A. Salama, G. Hassan, Attribute selection via a novel interval based evaluation algorithm: Applied on real life data sets, in: MATEC Web of Conferences, vol. 76, EDP Sciences, 2016, 04030.

[19] M. Durairaj, T.S. Poornappriya, Why feature selection in data mining is prominent? A survey, in: International Conference on Artificial Intelligence, Smart Grid and Smart City Applications, Springer, Cham, 2019, pp. 949–963.

[20] M.A. Hall, Correlation-based feature selection for machine learning (Ph.D. thesis), The University of Waikato, Hamilton, New Zealand, 1999.

[21] G. Manikandan, S. Abirami, Feature selection is important: State-of-the-Art Methods and application domains of feature selection on high-dimensional data, in: Applications in Ubiquitous Computing, Springer, Cham, pp. 177-196.

[22] M.B. Kursa, A. Jankowski, W.R. Rudnicki, Boruta–a system for feature selection, Fund. Inform. 101 (4) (2010) 271–285.

[23] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[24] M.B. Kursa, W.R. Rudnicki, The all relevant feature selection using random forest, 2011, arXiv preprint arXiv:1106.5112.

[25] A.J. Tallón-Ballesteros, J.C. Riquelme, R. Ruiz, Semi-wrapper feature subset selector for feed-forward neural networks: applications to binary and multi-class classification problems, Neurocomputing 353 (2019) 28–44.

[26] S.A. Medjahed, et al., Kernel-based learning and feature selection analysis for cancer diagnosis, Appl. Soft Comput. 51 (2017) 39–48.

[27] S. Jadhav, H. He, K. Jenkins, Information gain directed genetic algorithm wrapper feature selection for credit rating, Appl. Soft Comput. 69 (2018) 541–553.

[28] P.K. Theodoridis, D.C. Gkikas, Optimal feature selection for decision trees induction using a genetic algorithm wrapper-a model approach, in: Strategic Innovative Marketing and Tourism, Springer, Cham, 2020, pp. 583–591.

[29] K. Bache, M. Lichman, UCI Machine Learning Repository, 2014, University of California, Irvine, School of Information and Computer Sciences, Web Accessed(02/01/2014), http://archive.ics.uci.edu/ml.

[30] I. Guyon, et al., Result analysis of the NIPS 2003 feature selection challenge, Adv. Neural Inform. Process. Syst. (2005) 545–552.

[31] K. Wang, et al., Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer, Nature Genet. 46 (6) (2014) 573.

[32] R.A. Schiavo, D.J. Hand, Ten more years of error rate research, Internat. Statist. Rev. 68 (3) (2000) 295–310.

[33] D. Koller, M. Sahami, Toward Optimal Feature Selection, Stanford InfoLab, 1996.

[34] S. Shilaskar, A. Ghatol, Feature selection for medical diagnosis: Evaluation for cardiovascular diseases, Expert Syst. Appl. 40 (10) (2013) 4146–4153.

[35] X. Wang, K.K. Paliwal, Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition, Pattern Recognit. 36 (10) (2003) 2429–2439.

[36] S.-B. Cho, A.J. Tallón-Ballesteros, Visual tools to lecture data analytics and engineering, in: International Work-Conference on the Interplay Between Natural and Artificial Computation, Springer, Cham, 2017, pp. 551–558.

[37] E. Frank, et al., Weka-a machine learning workbench for data mining, in: Data Mining and Knowledge Discovery Handbook, Springer, Boston, MA, 2009, pp. 1269–1277.

[38] RS Team, RStudio: Integrated Development for R, RStudio, Inc., Boston, MA, 2015, URL http://www.rstudio.com.

[39] M.B. Kursa, W.R. Rudnicki, Feature selection with the boruta package, J. Stat. Softw. 36 (11) (2010) 1–13.

[40] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theory 13 (1) (1967) 21–27.

[41] E. Frank, I.H. Witten, Generating accurate rule sets without global optimization, 1998.

[42] P. Langley, W. Iba, K. Thompson, An analysis of Bayesian classifiers, Aaai 90 (1992) 223–228.

[43] A.J. Tallón-Ballesteros, J.C. Riquelme, Data mining methods applied to a digital forensics task for supervised machine learning, in: Computational Intelligence in Digital Forensics: Forensic Investigation and Applications, Springer, Cham, 2014, pp. 413–428.

[44] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Measur. 20 (1) (1960) 37–46.

[45] L. Yu, H. Liu, Feature selection for high-dimensional data: A fast correlation-based filter solution, in: Proceedings of the 20th international conference on machine learning, ICML-03, 2003, pp. 856-863.

[46] J.R. Quinlan, J. Ross, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.