Featuring the attributes in supervised machine learning

Antonio J. Tallón-Ballesteros[1], Luís Correia[2], and Bing Xue[3]

[1] Department of Languages and Computer Systems.
University of Seville, Seville, Spain
atallon@us.es
[2] BioISI - Faculdade de Ciências.
Universidade de Lisboa, Lisboa, Portugal
[3] School of Engineering and Computer Science.
Victoria University of Wellington, Wellington, New Zealand

Abstract. This paper introduces an approach to feature subset selection which is able to characterise the attributes of a supervised machine learning problem into two categories: essential and important features. Additionally, the fusion of both kinds of features yields to an overcoming in the prediction task, where some measures such as accuracy and Receiver Operating Characteristic curve (ROC) have been reported. The test-bed is composed of eight binary and multi-class classification problems with up to five hundred of attributes. Several classification algorithms such as Ridor, PART, C4.5 and NBTree have been tested to assess the proposal.

1 Introduction

Supervised Machine Learning (SML) via classification requires that every object has a label associated [6]. Essentially, classification partitions the whole feature space (the space of all possible attribute value combinations) into different regions, one for each class. The properties involved in a classification procedure may not always be manageable, which is more prone to happen when their number is high. Removing some of them alleviates the load of the learning machine induction and might lead a more accurate classification model. Lesser useful attributes for classification are detected and discarded, which is the operation performed by an attribute selection procedure.

The objective of this paper is to propose a new approach to feature selection, splitting it into two sequential stages: selection of *essential* attributes and selection of *important* attributes from the set of non-essential ones. The merge of these two sets is used to train the classifier. To evaluate the approach we use it with four different classifiers, namely Ridor, PART, C4.5 and Naïve Bayes tree (NBTree). This allows to assess the influence of using trees, rules and/or probabilistic approaches for the attribute subset selection model proposed.

The remaining of this article is arranged as follows. Section 2 provides a brief overview of different concepts about feature selection. Section 3 details the

proposal. Section 4 describes the experimentation by means of the approach setting, problems and classifiers used. Then, Section 5 depicts the empirical results. Lastly, Section 6 draws some conclusions.

2 Feature selection

Attribute selection is a specially important process for mining big data. Doing feature selection before a learning algorithm is applied has numerous benefits. By eliminating a significant amount of attributes it becomes easier to train learning machines. The computational time of the induction is reduced and the resulting model will usually be simpler and easier to interpret. It is also frequently the case that simpler models generalise better. Therefore, a model employing fewer features is likely to perform better. This is a process to determine from the instance set which attributes are more relevant to predict or explain the data, and conversely which attributes are redundant or provide little information [11]. Finally, the identification of the most relevant attributes can be useful in its own right providing valuable information about the problem in hand.

Generally speaking, three types of approaches might be used for attribute selection [9]: a) *Filter methods*, which select the best individual attributes usually assuming they are independent given the class. In this case some statistical measure is used to assess the quality of the attributes; b) *Wrapper methods* that use a machine learning algorithm to select a sub-set of the attributes. Usually this involves selection and evaluation of different sub-sets under some accuracy measure; c) *Embedded methods* combine the model creation problem with the attribute selection. These methods include in the induction model some bias towards fewer attributes.

By its part, the filter approaches may be divided into feature ranking and feature subset selection methods depending on the output which may be an ordered list of the attributes or a subset of attributes. This article focuses on filter method to obtain feature subsets. The main contribution of the current work is the ability to characterise groups of attributes into two types of categories, namely essential and important feature subsets.

3 Proposal

This paper proposes a way to categorise the features in supervised machine learning problems. According to our approach, there are two kinds of features: i) Essential features which represent the core properties to be collected from the new instances belonging to the problem; ii) Important features which constitute additional information that may be interesting to be reported on unseen instances. The procedure is as follows: a) first, the data set is divided into two sets: training and test sets, b) the feature selection method is applied to the training set and as an outcome we have the essential features which are those that have been picked up by the data preparation method and, on other bag, we have the non-selected properties that may not be thought to be very relevant

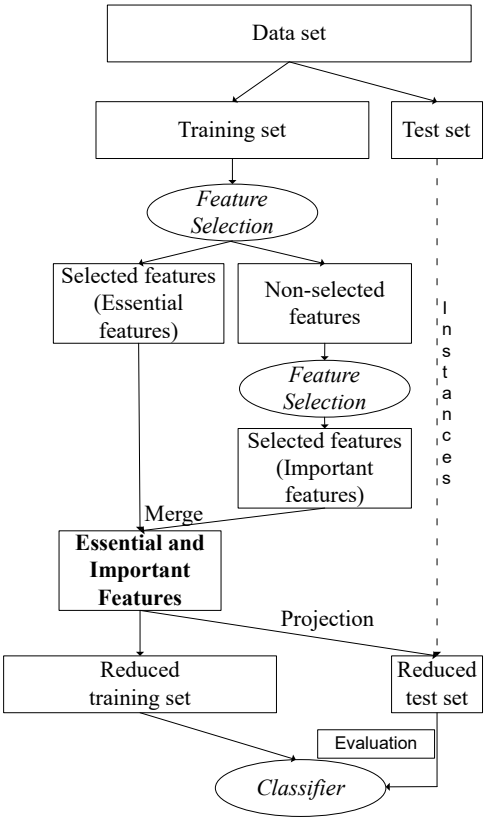

Fig. 1. Proposal: Characterisation of Features through Feature Subset Selection (ChaF2S2).

in terms of aid to the predictive data mining method, c) feature selection is performed on the non-selected attributes to extract the best from the not very promising features, d) the attributes from steps b) and c) are merged which will be the next characteristic space for classifier, e) the list of attributes is projected into the test set as it was originally and f) the usual assessment in data mining is conducted: we start training the classifier with the reduced training set and the evaluation takes place on the reduced test set. Figure 1 depicts the approach which has been named Characterisation of Features through Feature Subset Selection (ChaF2S2).

It is important to remark that some connections may be found with a previous work [21] due to the use of the merge operation. It is straightforward that in this paper no kind of overlapping [19] may occur, which may do the new approach more applicable and even more oriented to the goal that we have marked for the current contribution.

4 Experimentation

4.1 Approach setting

It is true that the amount of literature coping with many different measures is very extensive. We opt for correlation measures since the behaviour is very good and also has been one of the most commonly used by the data mining community. Moreover, our previous experience showed that the correlation is very convenient for supervised machine learning tasks [16]. Table 1 describes the methods to evaluate the current proposal which are founded on Correlation-based Feature Selection (CFS) [5] and Fast Correlation-Based Filter (FCBF) [25]. The reason of this chosen is motivated by the good performance of these feature subset selectors.

Specifically, we use the implementations provided by Weka tool [1] which are called CfsSubsetEval and FCBFSearch working with SymmetricalUncertAttributeSetEval, respectively for CFS and FCBF. CFS procedure has been used for FSS1 and FSS2 methods whereas FCBF has been utilised for FSS3 and FSS4 selectors. FSS1 and FSS3 capture the subset of essential features and FSS2 and FFS4 incorporate an extra subset of attributes, which we have called important features, to the solutions got by FSS1 and FSS3, respectively. FSS2 and FSS4 are the more complete options within their category and are the base of the current contribution. As an additional breakthrough, the distinction between essential and important features has been outlined. Table 2 reports on the parameters and properties to set up the method and also to ease the reproducibility of the experiments. It is important to remark that the experiments have been conducted with the default values parameters because the own authors have recommended them. Moreover, we also tested for CFS three deeper levels for the number of expanded nodes such as 6, 7 and 8; since there are not differences in the reached solutions we keep the number of expanded nodes to 5.

Table 1. Feature subset selectors for the experimentation

Abbreviation	Method	Essential features	Important features
$FSS1$	CFS	Yes	No
$FSS2$	CFS	Yes	Yes
$FSS3$	$FCBF$	Yes	No
$FSS4$	$FCBF$	Yes	Yes

4.2 Problems

A good range of problems have been tested to evaluate the performance of the new proposal. Table 3 summarises the test-bed. Their source is varied since some of them are available in the very well-known repository from the University of

Table 2. Parameter values and description of CFS and FCBF feature subset selectors

Method	Parameter/Property	Value
CFS	*Attribute evaluation measure*	*Correlation*
	Search method	*Best First*
	Consecutive expanded nodes without improving	5
	Search direction	*Forward*
FCBF	*Attribute evaluation measure*	*Correlation*
	Attribute evaluator	*Symmetrical Uncertainty*
	Search method	*FCBFSearch*

California (UC) at Irvine [22], MADELON has been proposed in NIPS 2003 challenge [4] and STAD is a Bioinformatics problem [23] that stands for STomach ADenocarcinoma. There are five multi-class problems and the remaining are binary, throwing there an average close to 4. The dimensionality goes from around 10 to 500 with an average over 82 whereas the data size fluctuates from one hundred to nineteen hundred. Nowadays, the number of attributes that we may have in problems at hand is very high and it is not strange to have thousands of features [18].

The data partition in some problems such as Led24 and SPECTF [8] follows the original pre-arrangement [14] and in most of the cases has been obtained with a stratified hold-out keeping the original data distribution in both sets, namely training and testing sets. Regarding the data imputation, a single imputation method called mean or mode imputation [13] has been applied which imputes a missing value with the mean or the mode within the class. We have adopted this strategy since the amount of missing values is very small. The data preparation method at the feature level has been only conducted to the training set and hence to get the reduced test only the projection operator is applied.

Table 3. Supervised machine learning problems

Problem	*Classes*	*Instances*			*Features*				
		Total	*Tra.*	*Tes.*	*Ori.*	*CFS*		*FCBF*	
						Ess.	*Imp.*	*Ess.*	*Imp.*
B. tissue	6	106	81	25	9	6	2	4	2
CTG	3	2126	1594	532	22	7	11	8	3
*Led*24	10	3200	200	3000	24	6	1	6	1
MADELON	2	2000	1500	500	500	12	4	7	2
Magic	2	19020	14265	4755	10	4	2	2	1
SPECTF	2	267	80	187	44	12	8	6	5
STAD	3	100	75	25	14	4	1	4	1
Waveform	3	5000	3750	1250	40	14	4	5	2
Average	3.9	3977.4	2693.1	1284.3	82.9	8.1	4.1	5.3	2.1

Tra. = Training Tes. = Testing Ori. = Original Ess. = Essential Imp. = Important

4.3 Classification algorithms

Different classifiers based on rules and trees have been used in this work, namely, Ridor and PART, from the former category, and C4.5 and NBTree, from the latter type, to assess how the new approach to feature subset selection performs in various conditions. We briefly review their characteristics. The proposal has been tested in two classic classifiers such as C4.5 and PART due to their good mixture with feature selection based on correlation as a previous contribution [20] to HAIS 2016 [10] reported. Ridor is an effective classification algorithm. Finally, NBTree is a very powerful classifier according to a very recent study in medium and high-dimensionality problems [15].

Ridor [3] is a ripple-down rule learner. It creates a default rule first and then the exceptions for the default rule with the least (weighted) error rate. Then it builds the best exceptions for each exception and iterates until pure. Thus it performs a tree-like expansion of exceptions. The exceptions are a set of rules that predict classes other than the default. PART [2] is a learning algorithm that generates a rule classifier using tree generation in the process. To generate a rule, a pruned decision tree is constructed for the current set of instances, then a rule is generated representing the leaf with the largest coverage, and the tree is discarded. The instances covered by the rule are removed and the process is repeated. The main advantage of PART is simplicity which allows it to scale with a high performance. C4.5 [12] builds a decision tree choosing for each node the attribute with the highest entropy. It can handle both discrete and continuous attributes. The implementation used also includes a pruning phase to reduce the tree structure and to improve the generalisation performance. The robustness and good interpretable results of this algorithm made it a popular choice in a variety of machine learning problems, and therefore we also chose it for this study. A NBTree classifier [7] generates a tree with Naïve Bayes (NB) classifiers at the leaves. To define each new node a univariate split is tested and the attribute with the highest utility is selected for that node. There is an exception when the utility is not significantly better than the utility of the current node, in which case a NB classifier is created for the current node. This model is as interpretable as trees and NB models while often showing better performance in large problems. It uses NB that is proven to be an optimal classifier under some circumstances [26] and it is usually taken as a reference classifier. This is the main reason for having chosen NBTree for testing our attribute selection model.

5 Results

This section compares every couple of related feature subset selection methods. Concretely, on the one hand, FSS1 and FSS2 are compared and, on the other hand, FSS3 faces FSS4. For the aforementioned classifiers, we report on the accuracy and Receiver Operating Characteristic curve (ROC) measure on the test set for each problem of the test-bed under all the scenarios described in 4.1.

5.1 Application of the proposal on Correlation-based Feature Selection (CFS)

Table 4 shows the test results of the proposal on CFS with Ridor classifier. The number of wins is higher than the losses. There are also some scenarios with ties. The proposal helps to enhance one or both assessment measures in most cases.

Table 5 depicts the performance of supervised machine learner PART. The scenario has completely changed from the previous classifier. Improvements have been reached in five out of the problems. Besides, the effect of the No-free lunch theorem is drawn around because only one measure is overcame in the half of the test-bed [24]; in particular, it happens improvement for B. tissue, CTG, Led24 and STAD in accuracy or ROC metric, exclusively.

Table 6 reports on the test results for classifier C4.5. In most of the data sets, it takes place improvement not only in accuracy but also in ROC. Moreover, there are two problems that may hint to be very difficult because it happen a worsening with both measures. STAD is a complex data set because: i) there are only 25 instances in the test set which means that every error in the prediction scores a negative 4%, ii) there are 3 classes and iii) is a Bio-informatics problem whose data have been collected very recently and the number of available measures is very low which makes the study a very challenging task. The results for STAD suggest that C4.5 is not a good option for this data set probably due to the cut-off values to create a decision node. SPECTF is a particular case because the important features may be discarded safely with no difference in performance; in addition, if we test with the data set without any kind of pre-processing the results are a bit better what suggests that feature selection may not be a good approach to deal with this problem [17].

Table 7 represents the behaviour of NBTree approach. The accuracy is enhanced in most cases, more concretely if the single tie is excluded, in five out of seven problems there is an overcoming. On the other way round, the ROC measure is often decreased what in any sense suggests to explore new ways or even to think about the option of only incorporating some of the important features. The good news here is that in two out of the top-3 problems in terms of features such as MADELON and Waveform a very noticeable progress has taken place.

5.2 Application of the proposal on Fast Correlation-Based Filter (FCBF)

Table 8 exhibits the performance via Ridor. Accuracy has been improved six times whereas ROC has been overcame four times. For those cases with negative outcomes the differences are very small which makes the approach very convenient and handy for the majority of the test-bed.

Table 9 shows the results with unseen data for the classification algorithm PART which is based on rules. There are many wins and only one or two losses for accuracy and ROC, respectively. In five out of the problems both measures are enhanced simultaneously which is very noticeable.

Table 4. Test results for the approach on CFS with Ridor

Problem	*Accuracy*			*ROC*		
	$FSS1$	$FSS2$	$Diff.$	$FSS1$	$FSS2$	$Diff.$
$B.Tissue$	60.00	56.00	−4.00	0.9000	0.9000	0.0000
CTG	78.20	80.64	2.44	0.7792	0.8560	0.0769
$Led24$	67.40	66.50	−0.90	0.8857	0.8275	−0.0582
$MADELON$	68.20	73.00	4.80	0.6820	0.7300	0.0480
$magic$	79.89	81.30	1.41	0.7782	0.7450	−0.0332
$SPECTF$	63.64	65.24	1.60	0.6806	0.6893	0.0087
$STAD$	64.00	64.00	0.00	0.6654	0.6654	0.0000
$Waveform$	76.88	76.72	−0.16	0.7552	0.7728	0.0176
$W/T/L$		4/1/3			4/2/2	

Table 5. Test results for the approach on CFS with PART

Problem	*Accuracy*			*ROC*		
	$FSS1$	$FSS2$	$Diff.$	$FSS1$	$FSS2$	$Diff.$
$B.\ tissue$	56.00	48.00	−8.00	0.9250	0.9300	0.0050
CTG	81.20	81.95	0.75	0.9190	0.8674	−0.0516
$Led24$	68.50	68.53	0.03	0.9227	0.9094	−0.0132
$MADELON$	60.80	62.60	1.80	0.7104	0.7295	0.0191
$Magic$	81.91	83.32	1.41	0.8712	0.8797	0.0086
$SPECTF$	70.05	72.19	2.14	0.6459	0.7000	0.0541
$STAD$	52.00	36.00	−16.00	0.5331	0.6581	0.1250
$Waveform$	77.04	76.80	−0.24	0.8432	0.8426	−0.0007
$W/T/L$		5/0/3			5/0/3	

Table 6. Test results for the approach on CFS with C4.5

Problem	*Accuracy*			*ROC*		
	$FSS1$	$FSS2$	$Diff.$	$FSS1$	$FSS2$	$Diff.$
$B.\ tissue$	68.00	56.00	−12.00	0.9250	0.8350	−0.0900
CTG	78.38	83.65	5.26	0.8967	0.9145	0.0177
$Led24$	68.10	68.80	0.70	0.8905	0.9079	0.0174
$MADELON$	70.60	73.60	3.00	0.7414	0.7826	0.0412
$Magic$	82.42	83.79	1.37	0.8653	0.8646	−0.0007
$SPECTF$	66.84	66.84	0.00	0.5519	0.5519	0.0000
$STAD$	72.00	52.00	−20.00	0.7574	0.6838	−0.0735
$Waveform$	74.40	76.16	1.76	0.7884	0.7879	−0.0005
$W/T/L$		5/1/2			3/1/4	

Table 7. Test results for the approach on CFS with NBTree

Problem	*Accuracy*			*ROC*		
	$FSS1$	$FSS2$	$Diff.$	$FSS1$	$FSS2$	$Diff.$
B. tissue	52.00	64.00	12.00	0.9250	0.9000	−0.0250
CTG	76.50	76.69	0.19	0.8413	0.7505	−0.0908
*Led*24	70.73	70.73	0.00	0.9685	0.9685	0.0000
$MADELON$	71.20	75.80	4.60	0.7693	0.8106	0.0413
$Magic$	81.93	83.11	1.18	0.8647	0.8747	0.0101
$SPECTF$	72.19	67.91	−4.28	0.7649	0.7103	−0.0547
$STAD$	64.00	56.00	−8.00	0.7096	0.6912	−0.0184
$Waveform$	76.88	81.36	4.48	0.8696	0.8916	0.0220
$W/T/L$		5/1/2			3/1/4	

Table 10 displays the behaviour of classifier C4.5. There are from 4 up to 5 wins according to the concrete metric and there is one tie. The situation for STAD problem has not been changed compared to the approach based on CFS; it seems that STAD may not be combined with a split criterion founded on entropy as C4.5 has.

Table 11 reports the test results for NBTree which is a tree-based approach built via the Bayes theorem. The outcome is very similar to the previous scenario although the differences for negative cases are smaller which leads to think that a probabilistic model is more suitable than traditional C4.5 algorithm, especially for STAD problem.

Table 8. Test results for the approach on FCBF with Ridor

Problem	*Accuracy*			*ROC*		
	$FSS3$	$FSS4$	$Diff.$	$FSS3$	$FSS4$	$Diff.$
B.Tissue	60.00	56.00	−4.00	0.9000	0.9000	0.0000
CTG	78.38	79.32	0.94	0.7804	0.7895	0.0091
*Led*24	67.37	68.37	1.00	0.8857	0.8569	−0.0289
$MADELON$	55.20	57.40	2.20	0.5520	0.5740	0.0220
$Magic$	77.60	81.47	3.87	0.6970	0.7558	0.0589
$SPECTF$	59.89	68.45	8.56	0.6907	0.6764	−0.0143
$STAD$	64.00	64.00	0.00	0.6654	0.6654	0.0000
$Waveform$	74.16	75.44	1.28	0.7489	0.7515	0.0026
$W/T/L$		6/1/1			4/2/2	

Once the results under two different scenarios have been depicted for the proposal, we must remark that CFS and FCBF are very good candidates to be used in future works although the performance of FCBF is stronger than CFS what may make the new approach an interesting option for data sets with a huge number of features.

Table 9. Test results for the approach on FCBF with PART

Problem	*Accuracy*			*ROC*		
	$FSS3$	$FSS4$	$Diff.$	$FSS3$	$FSS4$	$Diff.$
$B.Tissue$	48.00	48.00	0.00	0.9150	0.9300	0.0150
CTG	77.26	79.51	2.26	0.8909	0.9388	0.0479
$Led24$	68.50	68.90	0.40	0.9227	0.9373	0.0146
$MADELON$	60.40	60.40	0.00	0.6307	0.6227	−0.0080
$Magic$	79.26	82.54	3.28	0.8385	0.8648	0.0264
$SPECTF$	64.71	72.73	8.02	0.6983	0.6151	−0.0831
$STAD$	52.00	36.00	−16.00	0.5331	0.6581	0.1250
$Waveform$	74.00	74.24	0.24	0.8494	0.8561	0.0067
$W/T/L$		5/2/1			6/0/2	

Table 10. Test results for the approach on FCBF with C4.5

Problem	*Accuracy*			*ROC*		
	$FSS3$	$FSS4$	$Diff.$	$FSS3$	$FSS4$	$Diff.$
$B.Tissue$	48.00	56.00	8.00	0.8000	0.8350	0.0350
CTG	77.82	79.89	2.07	0.8546	0.9215	0.0668
$Led24$	68.10	68.10	0.00	0.8905	0.8905	0.0000
$MADELON$	58.60	59.20	0.60	0.6041	0.6097	0.0056
$Magic$	79.71	82.42	2.71	0.8174	0.8594	0.0420
$SPECTF$	67.91	66.84	−1.07	0.7116	0.6717	−0.0399
$STAD$	72.00	52.00	−20.00	0.7574	0.6838	−0.0735
$Waveform$	74.72	75.68	0.96	0.8636	0.8482	−0.0154
$W/T/L$		5/1/2			4/1/3	

Table 11. Test results for the approach on FCBF with NBTree

Problem	*Accuracy*			*ROC*		
	$FSS3$	$FSS4$	$Diff.$	$FSS3$	$FSS4$	$Diff.$
$B.Tissue$	48.00	52.00	4.00	0.9000	0.9350	0.0350
CTG	78.76	80.64	1.88	0.8384	0.8631	0.0247
$Led24$	70.73	70.73	0.00	0.9685	0.9685	0.0000
$MADELON$	61.20	61.00	−0.20	0.6393	0.6400	0.0007
$Magic$	80.13	82.73	2.61	0.8487	0.8731	0.0243
$SPECTF$	71.12	70.59	−0.53	0.7579	0.7083	−0.0496
$STAD$	64.00	56.00	−8.00	0.7096	0.6912	−0.0184
$Waveform$	75.60	79.20	3.60	0.8760	0.8912	0.0152
$W/T/L$		4/1/3			5/1/2	

6 Conclusions

This paper introduced a new approach to feature subset selection that is able to distinguish between essential and important attributes. Moreover, the combination of both types of features on CFS and FCBF feature subset selectors yielded to an enhanced performance of classifiers such as PART, Ridor -in an outstanding way-, C4.5 and NBTree compared to the selection of only essential attributes. The main idea achieved by this research is that there are some attributes which are crucial to have a good generalisation capacity; at the same time those attributes that seems not to be very promising are handy to be lead through a feature subset selection procedure in order to keep the best of the not so good potential attributes that may be called important features, which is the second best kind of attribute according our new approach. The empirical study was conducted on eight binary and multi-class problems from different areas and sources. The results revealed that some progress took place in terms of performance at the price of increasing a bit the characteristic space. Lastly, it must be mentioned that FCBF takes a greater advantage than CFS with the proposal. Nonetheless, the approach is also very convenient for CFS.

Acknowledgments This work has been partially subsidized by TIN2014-55894-C2-R project of the Spanish Inter-Ministerial Commission of Science and Technology (MICYT), FEDER funds, the P11-TIC-7528 project of the "Junta de Andalucía" (Spain) and by FCT, Portugal, under Grant UID/Multi/04046/2013.

References

1. R R Bouckaert, E Frank, M A Hall, G Holmes, B Pfahringer, P Reutemann, and I H Witten. Weka—experiences with a java open-source project. *The Journal of Machine Learning Research*, 11:2533–2541, 2010.
2. E Frank and I H Witten. Generating accurate rule sets without global optimization. In J. Shavlik, editor, *Fifteenth International Conference on Machine Learning*, pages 144–151. Morgan Kaufmann, 1998.
3. B R Gaines and P Compton. Induction of ripple-down rules applied to modeling large databases. *Journal of Intelligent Information Systems*, 5(3):211–228, 1995.
4. I Guyon, S Gunn, A Ben-Hur, and G Dror. Result analysis of the nips 2003 feature selection challenge. In *Advances in neural information processing systems*, pages 545–552, 2005.
5. M A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1999.
6. J Kacprzyk and W Pedrycz. *Springer handbook of computational intelligence*. Springer, 2015.
7. R Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, pages 202–207, 1996.
8. D Koller and M Sahami. Toward optimal feature selection. Technical report, Stanford InfoLab, 1996.
9. H Liu and H Motoda. *Feature extraction, construction and selection: A data mining perspective*, volume 453. Springer Science & Business Media, 1998.

10. F Martínez-Álvarez, A Troncoso, H Quintián, and E Corchado. *Hybrid Artificial Intelligent Systems: 11th International Conference, HAIS 2016, Seville, Spain, April 18-20, 2016, Proceedings*, volume 9648. Springer, 2016.
11. S Olafsson, X Li, and S Wu. Operations research and data mining. *European Journal of Operational Research*, 187(3):1429–1448, 2008.
12. J R Quinlan. *C4.5: programs for machine learning*, volume 1. Morgan kaufmann, 1993.
13. J L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.
14. P Somol, J Grim, and P Pudil. The problem of fragile feature subset preference in feature selection methods and a proposal of algorithmic workaround. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 4396–4399. IEEE, 2010.
15. A J Tallon-Ballesteros and L Correia. Medium and high-dimensionality attribute selection in bayes-type classifiers. In *Bioinspired Intelligence (IWOBI), 2017 International Work Conference on*, pages 121–126. IEEE, 2017.
16. A J Tallón-Ballesteros, C Hervás-Martínez, J C Riquelme, and R Ruiz. Improving the accuracy of a two-stage algorithm in evolutionary product unit neural networks for classification by means of feature selection. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 381–390. Springer, 2011.
17. A J Tallón-Ballesteros, C Hervás-Martínez, J C Riquelme, and R Ruiz. Feature selection to enhance a two-stage evolutionary algorithm in product unit neural networks for complex classification problems. *Neurocomputing*, 114:107–117, 2013.
18. A J Tallón-Ballesteros and A Ibiza-Granados. Simplifying pattern recognition problems via a scatter search algorithm. *International Journal for Computational Methods in Engineering Science and Mechanics*, 17(5-6):315–321, 2016.
19. A J Tallón-Ballesteros and 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*, pages 531–539. Springer, 2017.
20. A J Tallón-Ballesteros, J C Riquelme, and R Ruiz. Accuracy increase on evolving product unit neural networks via feature subset selection. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 136–148. Springer, 2016.
21. A J Tallón-Ballesteros, J C Riquelme, and R Ruiz. Merging subsets of attributes to improve a hybrid consistency-based filter: a case of study in product unit neural networks. *Connection Science*, 28(3):242–257, 2016.
22. ML UCI. Repository, the uc irvine machine learning repository, 2017.
23. K Wang, S T Yuen, J Xu, S P Lee, H HN Yan, S T Shi, H C Siu, S Deng, K M Chu, S Law, et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature genetics*, 46(6):573, 2014.
24. D H Wolpert. The supervised learning no-free-lunch theorems. In *Soft computing and industry*, pages 25–42. Springer, 2002.
25. L Yu and 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)*, pages 856–863, 2003.
26. H Zhang. The optimality of naive bayes. In Valerie Barr and Zdravko Markov, editors, *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*. AAAI Press, 2004.