On incremental wrapper-based attribute selection: experimental analysis of the relevance criteria

Pablo Bermejo, Jose A. Gámez, Jose M. Puerta
Intelligent Systems and Data Mining group
Computing Systems Department / $i^3\mathcal{A}$
Universidad de Castilla-La Mancha. Albacete, Spain
{pbermejo,jgamez,jpuerta,}@dsi.uclm.es

Abstract

This paper deals with the problem of feature subset selection in classification oriented datasets with a (very) large number of attributes. In such datasets the classical wrapper approaches become intractable due to the high number of wrapper evaluations to be carried out. One way to alleviate this problem is to use the so-called filter-wrapper approach, which consists in the construction of a ranking among the predictive attributes by using a filter measure, and then a wrapper approach is used by following the rank. In this way the number of wrapper evaluations is linear with the number of predictive attributes. The main contribution of this paper is the analysis of different relevance criteria used to decide when a new feature must be included or rejected in the selected subset. Experiments have been carried out with three different criteria and different strictness levels, and a statistical analysis is used to draw the conclusions about the best configurations to be used.

Keywords: Statistical test, feature selection, classification.

1 Introduction

Feature (or variable, or attribute) Subset Selection (FSS) is the process of identifying the input variables which are relevant to a particular learning (or data mining) problem [7, 5]. Though FSS is of interest in both supervised and unsupervised data mining, in this paper we focus on supervised learning, and concretely in the classification task. That is, we consider the existence of a distinguished variable (the class) whose value is known in the dataset instances. Classification oriented FSS carries out the task of removing most irrelevant and redundant features from the data with respect to the class. This process helps to improve the performance of the learnt models by:

- Alleviating the effect of the curse of dimensionality.
- Increasing the generalization power.
- Speeding up the learning and inference process.
- Improving model interpretability.

Besides, on the contrary of other reduction techniques (e.g. principal component analysis), FSS does not alter the original representation, so it preserves the original semantics of the variables, helping domain experts to acquire better understanding about their data by telling them which are the important features and how they are related to the class.

In supervised learning FSS algorithms can be (roughly) classified in three categories: (1) embedded methods; (2) filter methods; and, (3) wrapper methods. By embedded methods we refer to those algorithms, e.g. C4.5

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 638–645
Torremolinos (Málaga), June 22–27, 2008

[8] that implicitly use the subset of variables they need. Filter techniques are those that evaluate the goodness of an attribute or set of attributes by using only intrinsic properties of the data (e.g. statistical or information-based measures). Filter techniques have the advantage of being fast and general, in the sense that the resultant subset is not biased in favour of a concrete classifier. On the other hand wrapper algorithms are those that use a classifier (usually the one to be used later) in order to asses the quality of a given attribute subset. Wrapper algorithms have the advantage of achieving a greater accuracy than filters but with the disadvantage of being (by far) more time consuming and obtaining an attribute subset that is biased toward the used classifier.

During the last decade wrapper-based FSS has been an active area of research. Different search algorithms (greedy sequential, best-first search, evolutionary algorithms, etc.) have been used to guide the search process while some classifier (e.g. Naive Bayes, KNN, etc.) is used as a subrogate in order to evaluate the goodness of the subset proposed by the search algorithm. There is no doubt that the results provided by wrapper methods are better than those obtained by using filter algorithms, but the main problem is that they do not scale well. Thus, while datasets up to 100 or 500 variables were the standard in the last decade, with the venue of 2000's decade new datasets which involve thousands of variables appeared (e.g. genetics or information retrieval based datasets), and the result is that the use of pure wrapper algorithms is intractable in many cases [9].

With the idea of retaining the advantages of using a wrapper evaluation but avoiding to pay its high computational cost, a family of hybrid filter-wrapper algorithms has arisen [9, 3]. The idea is to use a filter measure in order to obtain a ranking of the attributes relevance with respect to the class. Then, a greedy algorithm is used to run over the ranking by adding those variables that are relevant to the classification process, where the relevance of including a new variable is measured in a wrapper way. The main advantage of this approach is that it retains a great part of wrapper advantages, while reducing the computational cost to $\mathcal{O}(n)$ wrapper evaluations instead of $\mathcal{O}(n^2)$ as happens with pure wrapper approaches (e.g. forward sequential), where n stands for the number of variables. When we deal with thousands of variables this point makes the difference between considering the task computationally feasible or not.

In this paper we deal with the above described rank-based filter-wrapper or incremental wrapper-based FSS method. Our contribution lies in analysing whether a new attribute must be considered relevant or not. In the simplest case a subset $\mathbf{S} \cup \{A\}$ is better than $\mathbf{S}$ if $acc_{\mathcal{C}}(\mathbf{S} \cup \{A\}) > acc_{\mathcal{C}}(\mathbf{S})$; that is, if the accuracy of classifier $\mathcal{C}$ when trained with the dataset projected over the subset $\mathbf{S} \cup \{A\}$ is strictly greater than the accuracy of the same classifier ($\mathcal{C}$) when trained with the dataset projected over the subset $\mathbf{S}$. Of course, this relevance criterion is quite sensitive to overfitting and usually introduces more features/attributes in the selected subset than those needed to obtain a similar degree of accuracy. Because of this, Ruiz et al. ([9]) propose to use a heuristic criterion based on a t-test ran over the output of a k-fold cross validation in order to assess if the improvement obtained when adding A to $\mathbf{S}$ is significant or not. The criterion is heuristic because of the small size of the sample (k=5) and the used confidence level ($\alpha = 0.1$). In this paper we experiment with different values for α and with different tests, trying to obtain conclusions about the impact of them over both the accuracy of the obtained classifier and the cardinality of the selected feature subset. The experiments are carried out over a suite of 7 microarrays-based datasets ranging from 2000 to 16000 predictive attributes.

The paper is organized in 5 sections apart from this introduction. In Section 2 we detail the incremental selection algorithm. Then, the relevance criteria to be analyzed are introduced in Section 3. Section 4 contains the design of the experiments, their results and an analysis of these. Finally, in Section 5 we

present our concluding remarks.

2 Incremental Wrapper-based FSS

In this Section we briefly revise the ranking-based filter-wrapper approach. We take as basis the BIRS (Best Incremental Ranked Subset) algorithm as introduced in [9] (fig. 1).

In	**T** training, M measure, $\mathcal{C}$ classifier
Out	**S**
1	list $R = \{\}$ // The ranking
2	for each attribute $A_i \in \mathbf{T}$
3	$Score$=$M_{\mathbf{T}}(A_i$, class)
4	insert A_i in R according to $Score$
5	BestAcc = 0
6	$\mathbf{S} = \emptyset$
7	for i = 1 to N // N = R.size()
8	$\mathbf{S}_{aux} = \mathbf{S} \cup R[i]$
9	AuxAcc = $\text{acc}_{\mathcal{C}}(\mathbf{S}_{aux}, \mathbf{T})$
10	if (AuxAcc $\triangleright$ BestAcc)
11	$\mathbf{S} = \mathbf{S}_{aux}$
12	BestAcc = AuxAcc

Figure 1: BIRS algorithm.

As we can see in BIRS algorithm, we mainly need three components:

• A measure to assess the dependence/correlation degree between each attribute and the class. In both [9] and [3] Symmetrical Uncertainty (SU) is used. SU is a nonlinear information theory-based measure that can be interpreted as a sort of Mutual Information normalized to interval [0,1]:

$$SU(A_i, C) = 2\left(\frac{H(C) - H(C|A_i)}{H(C) + H(A_i)}\right),$$

C being the class and $H()$ being the Shannon entropy. Attributes are ranked in increasing SU order; that is, more informative attributes are placed first.

• A way to evaluate the goodness of a proposed subset. By $acc_{\mathcal{C}}(\mathbf{S}_{aux}, \mathbf{T})$ we refer to the accuracy of classifier $\mathcal{C}$ when training by using the projection of **T** over $\mathbf{S}_{aux}$. In [9] the accuracy is measured by using a 5 fold cross validation. Thus, the accuracy over each one of the 5 partitions is returned and the average is used as accuracy for subset $\mathbf{S}_{aux}$.

• A relevance criterion in order to decide if a new attribute A_i must be (or not) included in **S**. In BIRS this criterion is divided in two steps. First, the accuracy $AuxAcc = avg(a_{f_1}, \ldots, a_{f_5})$ is computed from the accuracies returned by the wrapper evaluator for each fold (folders from 1 to 5). If AuxAcc $\leq$ BestAcc then A_i is rejected, otherwise the significance of the improvement must be tested. To do this, in BIRS a paired t-test, with null hypothesis ($H_0 : AuxAcc = BestAcc$) is carried out by using as input the 5 accuracies of the current and best subset. The test is paired because the compared values are the corresponding accuracies for the same ith folder with a different attributes subset.
In [9] a significance level $\alpha = 0.1$ is used. The reasons to select such a high α are: (1) the goal is to obtain a heuristic relevance criterion; and, (2) because of the small sample size (5), smaller α values directly drive to avoid the inclusion of any attribute once the first one has been included. The relevance criterion is represented in Figure 1 with symbol $\triangleright$.

Obviously BIRS carries out exactly N wrapper evaluations. This number can be reduced if we stop after l consecutive rejections as in [3]. Thus, l is a tunable lookahead parameter, e.g. if $l = 0$ we get the stronger stopping criterion and if $l \leq N$ we get BIRS (used in this work).

In [9] a very complete experimental study is carried out by considering three classifiers (Naive Bayes, C4.5 and IB1) and several microarray datasets. After obtaining the selected subset **S** by using BIRS the accuracy for each dataset is computed by running a standard 10-folds cross validation using the projection of **S** over **T** and getting the mean accuracy. As a result, BIRS has (at least) a similar performance with respect to accuracy (i.e. there is no statistically significant difference) and selects a significantly smaller subset of attributes, with respect to some

state of the art FSS algorithms (sequential forward selection (SF), FOCUS, Correlation based FSS (CFS) and Fast Correlation Based Filter (FCBF)[12]). In fact, SF and CFS do not produce results for the largest datasets due to CPU time requirements.

3 Relevance criteria comparison

The study carried out in [9] is based on the comparison of BIRS with different FSS algorithms over several datasets and using different classifiers. However, in all the cases the same relevance criterion is used. The criterion used in BIRS has as main characteristic the one of being heuristic but based on an objective criterion (the output of a statistical test). In this work we plan to study alternatives to this relevance criterion by analysing: (1) the impact of the confidence level (α) in relation with the number of selected variables; (2) the use of a non-parametric test instead of a parametric one; and (3) an alternative significance criterion. As mentioned before, the wrapper evaluator carries out a k-fold cross validation over the training set **T**. Then, for each fold f_i a partition into training (f_{t_i}) and validation (f_{v_i}) set is done. The classifier is trained with f_{t_i} and evaluated by using f_{v_i} in order to obtain the corresponding accuracy a_{f_i}. Thus, the inputs for all the relevance criterion here considered are the sets of values $\mathbf{Acc} = \{a_{f_i}, \ldots, a_{f_k}\}$ and $\mathbf{Acc}^b = \{a^b_{f_i}, \ldots, a^b_{f_k}\}$ which corresponds to the accuracies returned by the wrapper evaluator when executed over the current feature set and over the best feature set seen. Let $AuxAcc$ and $BestAcc$ be the average accuracies for **Acc** and $\mathbf{Acc}^b$ respectively, then, we assume that the following criteria are applied in order to study the inclusion of the proposed attribute only if $AuxAcc > BestAcc$, otherwise it is directly discarded. Below we describe the three basic relevance criteria studied in this paper:

Criterion 1. Student T-test.- We first considered the relevance criterion proposed in [9]; that is, a parametric statistical test: paired Student's test. This is likely one of the most used test in machine learning statistical analysis, however it assumes a Gaussian distribution over the paired differences between the two datasets, which is not always satisfied. This is the case here because of the small sample size ($k = 5$), but as Ruiz et al. [9] point out the goal is to have an objective criterion about the relevance of including a new feature, not to make a statistical analysis of the populations. Another known problem of this test is that it is affected by outliers. In [9] the authors set $\alpha = 0.1$, in this paper our goal is to study the effect of using less restrictive α values (0.1, 0.15, 0.2 and 0.25).

Criterion 2. Wilcoxon signed-ranks test [10].- With the same idea of using an objective criterion as the described in the previous paragraph but avoiding the Gaussianity assumption, we study the use of a non-parametric test. In this case we select Wilcoxon signed-ranks test because it is one of the most frequently used in the machine learning literature. In [1] an expression to compute the z statistic value is provided for large sample size cases (e.g. > 25) but, since this is not our case, we have used exact z statistic values (which can be found in many statistics books) for our determined alpha values and samples size. This test is not so affected by outliers (as the t-test) since it checks values of paired differences instead of values of each sample. This is a rather important advantage specially in the case of having really few samples. As in the previous case we experiment with α=0.1, 0.15, 0.2 and 0.25.

Criterion 3. Minimum better folds heuristic. In this case we try with a pure heuristic criterion that tries to reject the same null hypothesis than in the previous cases in favor of the same alternative hypothesis, e.g., the mean of the values in set **Acc** is significantly different to mean of values in set $\mathbf{Acc^b}$. Thus, with the idea of avoiding to include a new feature because a noisy result, we impose that apart from $AuxAcc > BestAcc$, it must hold $a_{f_i} > a^b_{f_i}$ at least in min folds. The value of min plays the role of α in this

criterion. We try with min=2, 3, 4 and 5. The value of $min = 1$ is not considered because it favours the influence of outliers.

Initially, $\alpha = 0.05$ was also tried, but because the small sample size it turns in a really strict criterion so its results are not included.

4 Experiments

In this Section we run the BIRS algorithm with the different criteria stated in the previous Section. The accuracy of the resulting models (i.e. running the classifier over the selected subset) is measured by using a 10 folds cross-validation. With respect to the classifier we only consider Naive Bayes (NB) [2], which is quite sensitive to the set of attributes used as input. Concretely we have used WEKA [11] implementation of NB which models numerical variables by using uni-dimensional Gaussian distributions.

4.1 Datasets

We have run our experiments over 7 publicly obtained microarrays-based datasets, all of them related to cancer prediction. Datasets *Colon, Leukemia, Lymphoma* and *GCM* are the same used in [9] and can be downloaded in .arff format (e.g. WEKA data mining suite input format from site `http://www.upo.es/eps/aguilar/datasets.html`. Datasets *DLBCL-Stanford, ProstateCancer* and *LungCancer-Harvard2* can be downloaded from site `http://sdmc.i2r.a-star.edu.sg/rp/`.
Table 1 shows the number of features and records each dataset contains and also the accuracy achieved for each one when running a 10cv by using NB classifier. The last row shows the mean values for each column.

As it can be seen in Table 1 some datasets have a very high dimensionality so a fine reduction without decreasing their accuracy (or even improving it) might be very important for their processing in prediction tasks, due to both the reduction of CPU time and the gain insight on the knowledge of relevant features (genes) with respect to the studied cancer.

Table 1: Properties of the data sets.

Dataset	**#Features**	**Size**	**Acc.(%)**
Colon	2000	62	53.23
Leukemia	7129	72	98.61
Lymphoma	4026	96	75.00
GCM	16063	190	66.84
DLBCL	4026	47	97.87
Prostate	12600	136	55.88
Lung	12533	181	98.34
Mean	8340	112	77.97

4.2 Experiment design and results

The design of the experiments is easy, we simply run BIRS by using each one of the proposed criteria (3 criteria $\times$ 4 α values = 12 criteria) over each one of the 7 datasets. The results are shown in Tables 3, 4 and 5 (we use #features to refer to the mean number of features selected over the performed 10 folds cross-validation). However, in order to have a baseline results for the analysis of balance between number of selected features and obtained accuracy, we first run BIRS by using a greedy relevance criterion, i.e., $AuxAcc > BestAcc$?. The results of this algorithm called SimpleBIRS are shown in Table 2.

4.3 Comparison of criteria

From the tables, the first comment can be that the more strict is the significance level (α or min) used, the fewer the number of variables included in the selected subset. Although more observations of this type can be drawn, in order to back our conclusions, we have carried out the statistical analysis described below. Because we have multiple algorithms (criteria) and multiple datasets we follow the recommendations in [1] and run the Friedman test [4] followed by a post-hoc Holm test [6]. Friedman test is used for statistical comparison over three or more sets of values; in our case the inputs of the study are the set of mean accuracies (and mean number of features selected) computed for each microarray in each one of the 10 folds. The application of Friedman test only decides if there exists at least one set of values (e.g. one algorithm)

Table 2: Results for SimpleBIRS

Dataset	**#Features**	**Acc.(%)**
Colon	6.3	79.03
Leukemia	3.7	93.06
Lymphoma	11.7	77.08
DLBCL	3.7	91.49
Prostate	12.2	78.68
Lung	3.9	98.90
GCM	50.8	64.74
Mean	13.2	83.28

Table 3: Results when considering T-test as relevance criterion.

Dataset	α =**0.1**		α =**0.15**		α =**0.2**		α =**0.25**	
	Acc.	**#f**	**Acc**	**#f**	**Acc.**	**#f**	**Acc.**	**#f**
Colon	82.26	2.4	77.42	2.6	77.42	3.1	82.26	3.4
Leukemia	86.11	1.6	86.11	1.6	90.28	2.2	90.28	2.2
Lymphoma	67.71	5.2	63.54	5.1	73.96	7.2	72.92	7.5
DLBCL	87.23	1.6	87.23	1.6	85.11	1.8	87.23	1.8
Prostate	74.26	4.1	75.74	4.7	75.00	6.8	75.00	6.6
Lung	96.13	1.6	96.13	1.6	97.24	2.4	97.24	2.4
GCM	54.21	12.4	60.53	13.3	59.47	19.4	60.53	18.5
Mean	78.27	4.1	78.10	4.4	79.78	6.1	80.78	6.1

Table 4: Results when considering signed rank test as relevance criterion.

Dataset	α =**0.1**		α =**0.15**		α =**0.2**		α =**0.25**	
	Acc.	**#f**	**Acc**	**#f**	**Acc.**	**#f**	**Acc.**	**#f**
Colon	82.26	2.1	77.42	2.6	79.03	2.7	79.03	2.7
Leukemia	84.72	1.2	86.11	1.6	84.72	1.6	84.72	1.6
Lymphoma	69.79	3.9	63.54	5.2	64.58	5.3	64.58	5.3
DLBCL	80.85	1.3	87.23	1.6	87.23	1.5	87.23	1.5
Prostate	75.74	3.3	77.21	4.4	79.41	4.8	79.41	4.8
Lung	96.13	1.1	96.13	1.1	96.13	1.1	96.13	1.1
GCM	51.58	9.4	53.16	12.1	58.95	14.1	58.95	14.1
Mean	77.30	3.2	77.26	4.2	78.58	4.5	78.58	4.5

Table 5: Results when considering min folds better as relevance criterion.

Dataset	**min=2**		**min=3**		**min=4**		**min=5**	
	Acc.	**#f**	**Acc**	**#f**	**Acc.**	**#f**	**Acc.**	**#f**
Colon	80.65	3.8	80.65	3.0	83.87	2.2	74.19	1.9
Leukemia	87.50	2.5	86.11	1.7	83.33	1.2	83.33	1.1
Lymphoma	76.04	8.8	71.88	6.1	65.63	4.1	66.67	3.2
DLBCL	85.11	1.9	87.23	1.5	80.85	1.3	78.72	1.1
Prostate	77.94	11.1	79.41	7.2	75.74	3.7	75.74	2.6
Lung	97.24	2.7	96.13	1.7	96.69	1.2	96.13	1.0
GCM	64.21	36.6	64.74	24.5	50.53	11.4	47.37	5.8
Mean	81.24	9.6	80.88	6.5	76.66	3.6	74.59	2.4

which is different to at least another set of values (algorithm). Once we know that is the case, we run the post-hoc Holm test by choosing a control set of values and then comparing it with the rest of sets. Our comparison process is performed as follows:

1. First, we try to identify for each relevance criterion (i.e. t-test, Wilcoxon and min better folds) the best significance (α) values. To do this, we perform a Friedman test for each criterion regarding only accuracies as inputs and using also as input the results provided by SimpleBIRS (that is our baseline algorithm). So, in this step we run three Friedman test (one per criterion) in order to identify those configurations ($\langle$criterion,significance-value$\rangle$) that are not statistically different from the accuracy achieved by SimpleBIRS. As in two of the three cases Friedman test returns that in fact there is at least one (statistically significant) different algorithm, we run the post-hoc Holm test by choosing as control the set of values provided by SimpleBIRS. Table 6 show the results of this process, where cells crossed out means that they are significantly different from SimpleBIRS (the control) by using Holm test (p-value < 0.05). Thus, these algorithms are ruled out and therefore not considered in the subsequent steps.

Table 6: Results of step one: local comparison for each relevance criterion with respect to accuracy.

α=	**0.1**	**0.15**	**0.2**	**0.25**
min=	**2**	**3**	**4**	**5**
T-test	~~78.3~~	~~78.1~~	~~79.8~~	80.8
Wilcoxon	77.3	77.3	78.6	78.6
MinValues	81.2	80.9	~~76.7~~	~~74.6~~
SimpleBIRS	• 83.3			

2. In the second step we consider a single pool with all the survivor algorithms from the previous phase. Thus, we repeat the previous process over the eight algorithms (1 using T-test criterion, 4 using Wilcoxon criterion, 2 using min folds better criterion and SimpleBIRS). Table 7 shows the results, where we can see how two new algorithms are ruled out when considering this global analysis.

Table 7: Results of step two: global comparison with respect to accuracy.

α=	**0.1**	**0.15**	**0.2**	**0.25**
min=	**2**	**3**	**4**	**5**
T-test				80.8
Wilcoxon	~~77.3~~	~~77.3~~	78.6	78.6
MinValues	81.2	80.9		
SimpleBIRS	• 83.3			

3. To this point we have obtained a set of six algorithms such that the global statistical analysis does not find significant differences among them. Therefore, it is time to consider the number of selected variables by them. Thus, we repeat the previous process (Friedman + Holm) but taking as inputs the set of values related to the mean number of selected features for each microarray in each one of the 10 folds. The results are shown in Table 8 (notice that now the control is the algorithm with the smallest selected subset). We can observe that two configurations are ruled out, including the baseline algorithm.

Table 8: Results of step three: global comparison with respect to the number of selected features.

α=	**0.1**	**0.15**	**0.2**	**0.25**
min=	**2**	**3**	**4**	**5**
T-test				6.1
Wilcoxon			• 4.5	4.5
MinValues	~~9~~	6.5		
SimpleBIRS	~~1~~			

As we can see, from the originally 13 considered configurations (relevance criterion and strictness level), after the global statistical analysis, we have obtained a set of four configurations whose results are *non-significantly*

different neither with respect to accuracy nor with respect to the number of features. As it could be expected, neither maximum accuracy nor minimum number of selected features have been able to remain selected and only configurations with a good balance between accuracy and number of selected features have survived.

As pointed out in [9], incremental wrapper selection works better when using an objective relevance criterion (e.g. BIRS) than when using only an improvement in the mean accuracy as criterion (e.g. SimpleBIRS). However, we have detected that when using a T-test based relevance criteria, it is better to use a more relaxed confidence level (0.25 vs 0.1). The same conclusion about the strictness applies to the other two criteria (Wilcoxon and min folds better), therefore, we can conclude that a more relaxed confidence level allow the introduction of some extra feature which helps to improve the accuracy. Also of interest is to observe that using a non-parametric test is a clear alternative and even the use of a pure heuristic criterion which defends itself from noise and outliers by forcing to the selected subset to achieve an improvement in at least three (of the five) folds.

5 Conclusions

In this work we have carried out an experimental analysis about the relevance criterion used in incremental wrapper FSS selection algorithms. We have considered the T-test based criterion originally proposed in [9] and proposed another two more criteria. Besides, a study about the impact of the strictness level using in the relevance criterion has been also carried out. As a result we can conclude that our analysis corroborates the conclusion of Ruiz et al. [9] about the improvement of using an objective criterion when deciding if a new feature must be selected or not. However, our study points out that it is appropriate to relax the strictness level in order to get a better balance between accuracy and number of selected features. Furthermore, we have proved that apart of the T-test based criterion, the use of a non-parametric test (which avoids the normality assumption) and even of a purely heuristic criterion provides equivalent results. For the future we plan to propose new incremental wrapper algorithms by using min folds better as relevance criterion (min=3), because being equivalent in results to the other two criteria it is computationally cheaper.

Acknowledgements

This work has been supported by the JCCM under projects PBI-05-022/PCI-08-0048, MEC under projects TIN1504-06204-C03-03/TIN2007-67418-C03-01 and FEDER funds.

References

[1] J. Demsar. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7:1–30, 2006.

[2] R. Duda and P. Hart. Pattern classification and scene analysis. John Wiley and Sons, 1973.

[3] J. Flores and J. Gámez. Breeding value classification in manchego sheep: a study of attribute selection and construction. Lecture Notes in Computer Science, 3682:1338–1346, 2005.

[4] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32:675–701, 1937.

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

[6] S. Holm. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6:65–70, 1979.

[7] H. Liu and H. Motoda. Feature Extraction Construction and Selection: a data mining perspective. Kluwer Academic Publishers, 1998.

[8] J. Quinlan. Induction of decision trees. Machine Learning, 1:81–106, 1986.

[9] R. Ruiz, J. C. Riquelme, and J. S. Aguilar-Ruiz. Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recogn., 39:2383–2392, 2006.

[10] F. Wilcoxon. Individual comparisons by ranking methods. Biometrics Bulletin, 1:80–83, 1945.

[11] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2005.

[12] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 5:1205–1224, 2004.