Analysis of Feature Rankings for Classification

Roberto Ruiz, Jesús S. Aguilar–Ruiz,
José C. Riquelme, and Norberto Díaz–Díaz

Department of Computer Science, University of Seville, Spain
{rruiz, aguilar, riquelme, ndiaz}@lsi.us.es

Abstract. Different ways of contrast generated rankings by feature selection algorithms are presented in this paper, showing several possible interpretations, depending on the given approach to each study. We begin from the premise of no existence of only one ideal subset for all cases. The purpose of these kinds of algorithms is to reduce the data set to each first attributes without losing prediction against the original data set. In this paper we propose a method, *feature–ranking performance*, to compare different feature–ranking methods, based on the Area Under Feature Ranking Classification Performance Curve (AURC). Conclusions and trends taken from this paper propose support for the performance of learning tasks, where some ranking algorithms studied here operate.

1 Introduction

It is a fact that the performance of most practical classifiers improve when correlated or irrelevant features are removed. Feature selection attempts to select the minimally sized subset of features according to two criteria: classification accuracy does not significantly decrease; and resulting class distribution given only the values for the selected features, is as close as possible to the original class distribution, given all features. In general, the application of feature selection helps all phases of the data mining process for successful knowledge discovery.

Feature selection algorithms can be grouped into two categories from the point of view of a method's output: subset of features or ranking of features. One category is about choosing a minimum set of features that satisfies an evaluation criterion; the other is about ranking features according to same evaluation measure. Ideally, feature selection methods search through the subsets of features and try to find the best one among the competing 2^m candidate subsets (m: number of whole features), according to some evaluation function. However, this exhaustive process may be costly and practically prohibitive, even for a medium–sized feature set size. Other methods based on heuristic or random search methods attempt to reduce computational complexity by compromising performance.

When feature selection algorithms are applied as a pre–processing technique for classification, we are interested in those attributes that better classify new unseen data. If the feature selection algorithm provides a subset of attributes, this subset is used to generate the knowledge model that will classify the new

A.F. Famili et al. (Eds.): IDA 2005, LNCS 3646, pp. 362–372, 2005.

data. However, when the algorithm provides a ranking it is not easy to determine how many attributes are necessary to obtain a good classification result.

In this work, we present different ways to compare feature rankings and show the variety of possible interpretations depending on the study approach made. Our intent is to learn if any dependence between classifier and ranking methods exist as well as trying to answer two essential enquiries: What is a good feature ranking? And, how do we value/measure a ranking? To this end, we practise different comparisons using four feature ranking methods: χ^2, Information Gain, ReliefF and SOAP, which are commented on later. We will check the results by calculating the success rate using three classifiers: C4.5, Naïve Bayes and nearest neighbour.

The paper is organized as follows. In Section 2, concepts used throughout the paper are defined. Section 3 reviews related work and the motivation of our approach is presented, feature ranking methods and classification techniques to be used in the experiments are described. The AURC is shown in Section 4, experimental results in Section 5 and finally, in Section 6, the most interesting conclusions are summarized.

2 Definitions

In this section some definitions are given to formally describe the concepts used throughout the paper: feature ranking, classifier, classification accuracy and ranking–based classification accuracy.

Definition 1 (Data). *Let* D *be a set of* N *examples* $e_i = (\overline{x_i}, y_i)$*, where* $\overline{x_i} = (a_1, \ldots, a_m)$ *is a set of input attributes and* y_i *is the output attribute. Each input attribute belongs to the set of attributes* ($a_i \in$ A*, continuous or discrete) and each example belongs to the data* ($e_i \in$ D*). Let* C *be the decision attribute* ($y_i \in$ C*), named class, which will be used to classify the data. For simplicity in the paper,* y_i *means "the class label of the example* e_i*".*

Definition 2 (Feature Ranking). *Let* $\mathrm{A} = \{a_1, a_2, \ldots, a_m\}$ *be the set of* m *attributes. Let* r *be a function* $r : \mathrm{A_D} \rightarrow \mathbb{R}$ *that assigns a value of merit to each attribute* $a \in$ A *from* D*. A feature ranking is a function* F *that assigns a value of merit (relevance) to each attribute* ($a_i \in$ A*) and returns a list of attributes* ($a_i^* \in$ A*) ordered by its relevance, with* $i \in \{1, \ldots, m\}$*:*
$F(\{a_1, a_2 \ldots, a_m\}) =< a_1^*, a_2^*, \ldots, a_m^* >$ *where* $r(a_1^*) \geq r(a_2^*) \geq \ldots \geq r(a_m^*)$.

By convention, we assume that a high score is indicative of a relevant attribute and that attributes are sorted in decreasing order of $r(a^*)$. We consider ranking criteria defined for individual features, independently of the context of others, and we also limit ourselves to supervised learning criteria.

Definition 3 (Classification). *A classifier is a function* H *that assigns a class label to a new example:* $\mathrm{H} : \mathrm{A}^p \rightarrow \mathrm{C}$*, where* p *is the number of attributes to be used by the classifier,* $1 \leq p \leq m$*. The classification accuracy (CA) is the average success rate provided by the classifier* H *given a set of test examples,*

i.e., the averaged number of times that H *was able to predict the class of the test examples. Let* $\mathbf{x}$ *be a function that extracts the input attributes from the example* e, $\mathbf{x} : \mathrm{A}^m \times \mathrm{C} \rightarrow \mathrm{A}^m$. *For a test example* $e_i^* = (x_i, y_i)$, *if* $\mathrm{H}(\mathbf{x}(e_i^*)) = y_i$ *then* e_i^* *is correctly classified; otherwise misclassified.*

In this paper, to measure the performance of the classifiers only the leaving–one–out method will be used, because it is not dependent on randomness, like k–fold cross–validation or hold out. In the next expression, if $\mathrm{H}(\mathbf{x}(e_i)) = y_i$ then 1 is counted, otherwise 0. $CA = \frac{1}{N} \sum_{i=1}^{N} (\mathrm{H}(\mathbf{x}(e_i)) = y_i)$. As we are interested in rankings, the classification accuracy will be measured with respect to many different subsets of the ranking provided by some feature ranking methods.

Definition 4 (Ranking–based Classification). *Let* S_k^{F} *be a function that returns the subset of the first* k *attributes provided by the feature ranking method* F *(*$S_k^{\mathrm{F}} : \mathrm{A}^m \rightarrow \mathrm{A}^k$*). The ranking–based classification accuracy of* H *will be as follows:*

$$CA_k(\mathrm{F}, \mathrm{H}) = \frac{1}{N} \sum_{i=1}^{N} \left(\mathrm{H}(S_k^{\mathrm{F}}(\mathbf{x}(e_i))) = y_i\right)$$

Note that S_1^{F} is the first (best) attribute of the ranking provided by F; S_2^{F} are the first two attributes, and thus up to m.

3 Preliminary Study

3.1 Related Work

There are few specific bibliographies where feature ranking comparison is defined. Liu and Motoda [1] comments on the use of learning curves to demonstrate the effect of adding attributes when a list of ordered attributes is provided. There is a paper [2], in which attribute ranking by means of only one subgroup are compared, that one receiving the best classification from all the subgroups needed to obtain the learning curve. But, picking features whose importance is greater than a threshold value [3,4], is more simple and divulged. Irrelevant features (whose values are random) that are used as a threshold in the application of algorithm ranking are inserted in [5].

All the ranking comparison is based on calculate the rankings performance. Two measures currently exist to analyze this; by means of its accuracy or by the area under ROC (Receiver Operating Characteristics) curve [6]. A ROC curve A is said to dominate another ROC curve B if A is always above and to the left of B. In the cases where two ROC curves do not dominate each other in the whole range, or when the class distribution and error costs are unknown, the area under ROC curve (AUC) is a good "summary" for comparing these. So, a curve A dominates to another curve B if $AUC(A) > AUC(B)$, where AUC(A) and AUC(B) denotes the area under ROC curve A or B, respectively, in the ROC space. The main limitation of this measure lies in that it is only easily applicable to problems with two classes. For a problem with c classes,

ROC space is composed of $c*(c-1)$ dimensions. This fact makes the use of this techniques in problems with a considerable number of classes practically inviable and so, although this measure is better than the previous (based on accuracy), we will not use it. Remember that in this paper we intend to show how the user can choose the best possible method according to what the user is looking for, independently of type of the data set.

In all works of ranking comparison previously mentioned, the measure used to calculate the ranking performance is the exactness obtained by a classifier, with k first features list being different in how the threshold is fixed. This posed the following questions: What exactly is being evaluated, the ranking, or the method to select features? Is this correct? The value which is used in comparison depends on three agents: generated ranking, method of fixing the threshold and learning algorithm. The fact is that the classification model´s exactness can change substantially depending on the features taking part; therefore the way of choosing features seems more important than the order in which they are chosen. Consequentially, we can say that comparisons will be right, but not complete. Our suggestion is to directly value the ranking, without depending on the selection method.

3.2 Description of Methods

We have chosen four criteria to rank attributes (see [7] for review), all of them very different from each other. These feature–ranking methods are briefly described next: χ^2 (CH) was first introduced by Liu an Setiono [8] as a discretization method and later shown to be able to remove redundant and/or irrelevant continuous features; **Information Gain** (IG) is based on the information–theoretical concept of entropy, a measure of the uncertainty of a random variable; **Relief** (RL) algorithm uses an approach based on the nearest-neighbour algorithm to assign a relevance weight to each feature. Relief was originally introduced by Kira and Rendell [9] and later enhanced by Kononenko [10]. Each feature's weight reflects its ability to distinguish among the class values; **Soap** (Selection of Attributes by Projections) evaluation criterion [3] (SP) is based on a unique value called NLC (Number of Label Changes). It relates each attribute with the label used for classification. This value is calculated by projecting data set elements onto the respective axis of the attribute (ordering the examples by this attribute), then crossing the axis from the beginning to the greatest attribute value, and counting the NLC produced.

Once feature rankings are obtained, we check the results calculating the success rate using three classifiers. They are chosen as representatives of different types of classifiers: c4.5 [11] (c4) is a tool that summarizes training data in the form of a decision tree. Along with systems that induce logical rules, decision tree algorithms have proved popular in practice. This is due in part to their robustness and execution speed, and to the fact that explicit concept descriptions are produced, which users can interpret; The naive Bayes [12] (nb) algorithm represents knowledge in the form of probabilistic summaries. It employs a simplified version of Bayes formula to decide which class a novel instances belongs

to; Nearest-Neighbour [13] (nn) simply finds the stored instance closest (according to a Euclidean distance metric) to the instance to be classified (we will use only one neighbour, 1NN).

3.3 Motivation

Firstly, we observe the quality of the four feature-ranking methods in respect to the tree classifiers, we will use the Glass2 data set (214 examples, 9 attributes, 2 classes), since it is a representative case to discuss our motivation. Table 1 shows the rankings for χ^2, Information Gain, Relief and SOAP. For each feature–ranking method, the row *rk* presents the ranking of attributes and, under this row, the classification performance for C4.5, Naïve Bayes and the Nearest Neighbour technique (using only one neighbour), by using the number of attributes from the ranking indicated in the first row, under "Subset". Classification accuracies (using the 9 attributes) from C4.5, Naïve Bayes and 1–NN are very different: 75.5%, 62.0% and 77.3%, respectively. For example, the most relevant attribute for χ^2 and IG was 7, for Relief 3 and SOAP 1. Using only the attribute 7 (CH and IG), C4.5 produced a classification success of 73.6. However, the classification success with attribute 3 was 57.7 (RL) and 77.3 with attribute 1 (SP). The second attribute selected by χ^2 and IG was 1, Relief selected 6 and SOAP, 7. The first three attributes for χ^2, IG and SOAP were the same, so these three classification results are equal. The fourth attribute breaks the tie. Several interesting conclusions can be drawn from the analysis of Table 1: (a) The four feature–ranking methods provide different rankings, what obvi-

Table 1. Feature–rankings for Glass2. FR: Feature–Ranking method (CH: χ^2; IG: Information Gain; RL: Relief; SP: Soap); Cl: Classifier (c4: C4.5; nb: Naïve Bayes; nn: 1–Nearest Neighbour); and rk: ranking of attributes.

		Subset								
FR	Cl	1	2	3	4	5	6	7	8	9
CH	rk	7	1	4	6	3	2	9	8	5
	c4	**73.6**	77.9	**82.2**	78.5	75.5	74.8	73.6	76.1	75.5
	nb	57.1	57.1	66.9	69.9	63.8	63.8	63.2	62.0	62.0
	nn	66.9	**79.7**	75.5	**82.8**	**88.3**	**81.0**	**77.9**	**77.9**	**77.3**
IG	rk	7	1	4	3	6	2	9	8	5
	c4	**73.6**	77.9	**82.2**	82.2	75.5	74.8	73.6	76.1	75.5
	nb	57.1	57.1	66.9	63.8	63.8	63.8	63.2	62.0	62.0
	nn	66.9	**79.7**	75.5	**84.7**	**88.3**	**81.0**	**77.9**	**77.9**	**77.3**
RL	rk	3	6	4	7	1	5	2	8	9
	c4	57.7	67.5	80.4	76.7	75.5	75.5	74.8	77.9	75.5
	nb	**62.0**	62.6	65.0	64.4	63.8	63.8	63.8	62.6	62.0
	nn	58.9	**75.5**	**81.0**	**83.4**	**88.3**	**83.4**	**81.6**	**81.6**	**77.3**
SP	rk	1	7	4	5	2	3	6	9	8
	c4	**77.3**	77.9	**82.8**	**81.6**	**81.6**	**84.1**	74.9	73.0	75.5
	nb	52.2	57.1	66.9	65.6	62.6	63.2	63.8	62.0	62.0
	nn	72.4	**79.7**	75.5	79.8	80.4	82.2	**81.6**	**77.3**	**77.3**

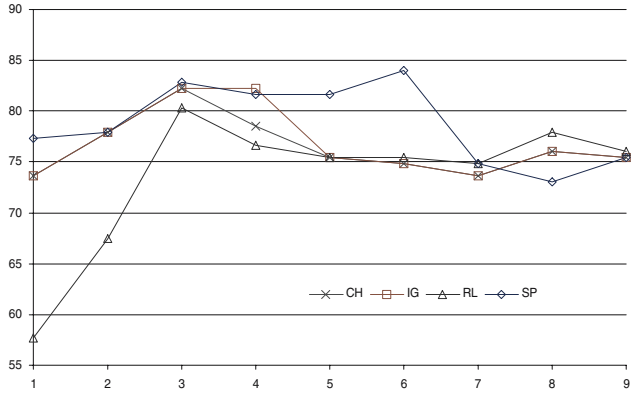

Fig. 1. Accuracy obtained by C4.5 for data set Glass2 (data from Table 1). The number of attributes used to classify are in the abscissa and the success rate in the ordinate.

ously leads to different classification performance. (b) The pair Soap+C4.5 is the only one that provides a classification performance (77.3) using only one attribute (attribute 1) better than using the whole set of attributes (75.5). (c) The sequence of best classification performance is, in principle, arbitrary: (SP+C4, 77.3), (SP+NN, 79.8), (SP+C4, 82.8), (IG+NN, 84.7), ({CH,IG,RL}+NN, 88.3), (SP+NN, 84.1), ({RL+SP}+NN,81.6), (RL+NN,81.6) and the last best value 77.3, with NN. (d) It seems that NN performs very well when the number of attributes is greater than $m/2$. A significant fact is that the best five attributes with 1NN are {1,3,4,6,7}, but the best six attributes are {1,2,3,4,5,7}. Attribute 6 is not that relevant when attributes 2 and 5 are taken into account. In general, a variable that is completely useless by itself can provide a significant performance improvement when it is taken with others.

Figure 1 shows the classification accuracy for C4.5 by using the four feature–ranking methods with the data set Glass2. Although the best subset exactness is similar, SOAP performance is excellent for any feature number and is the only method that in almost all subsets appears above average. In conclusion, we could assert that it is the best ranking of all. The analysis based on the best subset does not exactly show the kindness of features ranking because before or after that subset, the results could be terrible. Taking into account these conclusions, we want to consider the possibility of finding some insight about when one feature–ranking is better than others for a given classifier. Therefore, it would be interesting to explore the ranking method performance along the learning curve described, and extracting conclusions according to the feature proportion used.

Figure 2 shows the possible situations when we compare different rankings for a data subsets. The question posed is: Which ranking is better to classify? The answer would be conditioned by what the user is looking for. This means, if the interest is the ranking identification method that gets the best classified

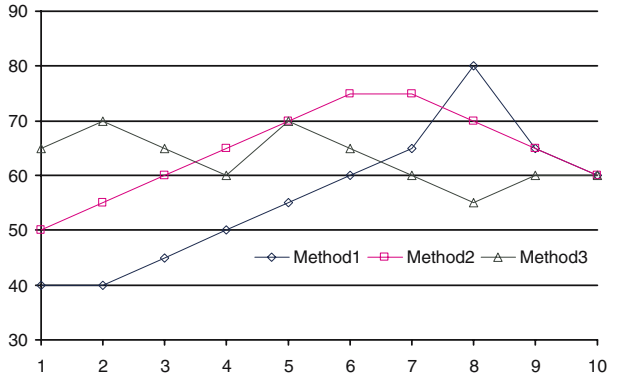

Fig. 2. Fictitious example of three different kind of learning curves

subset for a learning algorithm given, we should choose Method1, remaining conscious of what we need for that eighty percent of features. However, it has been observed that the rest of the curve classification results are almost always under two other methods. If we choose a features number lower than seventy percent, Method1 results will be the worst of the three. If what we are looking for is a best performance method along the whole curve, we must compare the evolution of the three curves point to point. Method2 loses at the beginning (until thirty percent of all features). With Method3, the former is always better than the previous ones, except in the previously commented case (with eighty percent of features). Finally, Method3 is the best, if we want to choose less than thirty percent of the features.

4 Area Under Learning Curve

Comparing subset to subset would be a more complete comparison between two features ranking. Comparing classification results obtained by the first feature of the two lists (the best one), with the two best, and so on successively until m ranked features. We could use this comparison, calculating the average of the obtained results with each list, to compare rankings. The calculation of the area under curve described by previous results would be a very similar study.

Area Under the Curve (AUC) is calculated applying the trapezium formula. In our case, the curve (learning curve) is obtained adding features according to the order assigned by ranking method.

$$\sum_{i=1}^{m-1} (x_{i+1} - x_i) * \frac{(y_{i+1} + y_i)}{2}$$

Definition 5 (AURC). *Given a feature ranking method* F *and a classifier* H*, we can obtain the performance of the classification method regarding the ranking*

provided by the feature ranking by measuring the area under the curve. The curve is drawn by joining every two points $(CA_k(\mathrm{F},\mathrm{H}), CA_{k+1}(\mathrm{F},\mathrm{H}))$*, where* $k \in \{1,\ldots,m-1\}$ *and* m *is the number of attributes. The Area Under Ranking Classification Performance Curve* $AURC(\mathrm{F},\mathrm{H})$ *will be calculated as:*

$$AURC(\mathrm{F},\mathrm{H}) = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} (CA_i(\mathrm{F},\mathrm{H}) + CA_{i+1}(\mathrm{F},\mathrm{H}))$$

With this expression, for any pair (F,H), $AURC(\mathrm{F},\mathrm{H}) \in [0,1]$ (in Table 3, it appears multiplied by one hundred for a better understanding), which provides us an excellent method to compare the quality of feature rankings with respect to each classification method. Take into account that the best AURC correspond to the best Ranking method.

An interesting property of this curve is that it is not monotone increasing, i.e., for some i, it would be possible that $CA_i(\mathrm{F},\mathrm{H}) > CA_{i+1}(\mathrm{F},\mathrm{H})$.

Definition 6 (Feature–Ranking Performance). *The feature–ranking performance is measured as the evolution of the AURC along the ranking of features, with step* $\delta\%$*. The curve is plotted, for every* $\delta\%$ *of the attributes as follows:*

$$AURC_\delta(\mathrm{F},\mathrm{H}) = \frac{1}{2\delta(m-1)} \sum_{i=1}^{\delta(m-1)} (CA_i(\mathrm{F},\mathrm{H}) + CA_{i+1}(\mathrm{F},\mathrm{H}))$$

We must consider that the idea concerning every feature selection method (one of them is ranking method) is that it must take the smallest number of features as possible. If we contemplate the possibility that in each learning curve, high and short exactnesses are compensated to the AURC calculation, we must make a study about methods performance using first features and fixed percentages.

5 Experiments

The implementation of induction algorithms and other selectors was done using Weka library [14] and comparison was performed with sixteen data sets from the University of California at Irvine [15] summarized in Table 2. All the experiments were run using leaving one out. The four methods of feature rankings are applied to each data set, and each ranking learning curve is calculated with the three classifiers.

Table 3 shows, for each data set, the Area Under Classification Performance Curve. Boldprint values are the best for the three classifiers, and those underlined are the best for corresponding classifiers. A clear conclusion can not be made, but specific trends can: (a) Results are very similar under each classifier (last line). There are some differences between each one of them. 1–NN is the classifier that offers a better performance with the four feature ranking methods; C4.5 is very close and NB is the last one. (b) If we take into account the best AURC for each data set, 1–NN obtains better results. (c) Most of the RL cases win, so that we could conclude that it is the best ranking method.

Table 2. Data sets used in the experiments

Data set	Id	Instances	Attributes	Classes
anneal	AN	898	38	6
balance	BA	625	4	3
g_credit	GC	1000	20	2
diabetes	DI	768	8	2
glass	GL	214	9	7
glass2	G2	163	9	2
heart–s	HS	270	13	2
ionosphere	IO	351	34	2
iris	IR	150	4	3
kr–vs–kp	KR	3196	36	6
lymphography	LY	148	18	4
segment	SE	2310	19	7
sonar	SO	208	60	2
vehicle	VE	846	18	4
vowel	VW	990	13	11
zoo	ZO	101	16	7

Table 3. AURC value for each ranking–classifier combination

	C4.5				NB				1NN			
DS	CHI2	IG	RLF	SOAP	CHI2	IG	RLF	SOAP	CHI2	IG	RLF	SOAP
an	97.30	97.12	96.90	97.11	85.82	86.30	86.50	86.47	**98.20**	98.09	97.54	97.71
bs	68.61	68.61	68.83	68.61	**75.55**	**75.55**	72.56	**75.55**	72.77	72.77	69.79	72.77
gc	72.39	72.39	71.71	72.31	**74.74**	**74.74**	73.89	74.22	70.16	70.16	70.38	66.83
di	72.85	72.89	73.30	72.52	75.36	**75.73**	75.68	75.15	68.87	69.52	68.09	67.87
gl	64.57	66.09	67.09	67.32	49.15	49.85	47.34	51.37	63.49	67.67	68.17	**71.12**
g2	76.65	77.11	74.35	79.03	63.27	62.50	63.50	62.27	79.41	79.64	**80.37**	78.91
hs	78.23	78.23	77.04	76.54	**83.09**	**83.09**	81.53	80.80	78.43	78.43	75.94	74.34
io	88.69	89.18	**90.03**	86.94	84.52	85.11	85..66	80.23	88.57	88.36	88.57	86.92
ir	95.11	95.11	95.22	95.22	95.56	95.56	95.56	95.56	95.00	95.00	95.56	95.56
kr	95.22	95.13	**96.48**	95.56	87.47	87.47	89..81	86.99	93.97	93.89	95.79	94.63
ly	74.84	75.97	75.66	75.42	78.76	80.25	80.37	80.09	76.61	81.28	**82.31**	80.56
se	92.23	92.15	93.37	92.96	74.63	73.79	78.26	76.77	92.78	93.11	**93.87**	93.26
so	73.92	73.71	76.06	75.15	67.62	67.44	69.57	68.83	84.15	83.94	**84.41**	83.87
ve	64.59	65.79	**68.10**	67.43	41.65	41.54	41.72	41.04	66.09	65.83	65.79	65.69
vw	73.98	74.20	73.96	74.59	61.96	62.46	61.65	62.17	**90.67**	90.66	89.63	90.52
zo	88.18	87.56	86.88	88.27	88.95	88.21	86.42	89.36	**91.34**	90.84	87.69	90.87
Av	79.83	80.08	80.31	80.31	74.25	74.35	74.37	74.18	81.91	82.45	**82.12**	81.96

In order to facilitate the comparison of diverse ranking methods from different points of view, and to extract some conclusions, table 4 is presented. In this table we show a summary of each time a ranking method holds the first position. Different groups of comparisons are set: results obtained by the first features

Table 4. Summary of times each ranking method holds first position. Results associated by: first features, percentages and classifiers

Results for	CH	IG	RL	SP
Exactness-1at:	26	28	20	**30**
AURC-2at:	18	22	16	**24**
AURC-3at:	15	17	15	**21**
AURC-4at:	14	15	16	**20**
AURC-5at:	15	**20**	18	17
AURC-25%:	17	**22**	21	17
AURC-50%at:	13	12	**21**	13
AURC-all at:	14	12	**25**	12
C4.5:	39	46	50	**51**
NB:	41	**59**	58	46
NN:	50	43	44	**55**
Total:	130	148	**152**	**152**

are situated in the first block (success rate with the first feature and the AURC with two, three, four and five features are contrasted); the second block shows the comparisons by percentage results (25, 50 with the whole results); and last group is broken down by classifiers.

If we contemplate the tests done by the first features, SOAP ranking method stands out, especially in relation with C4.5 and 1–NN classifiers, the one that offers the best result with NB is IG, using only the first ranking features. IG and RL obtain better results at 25% of ranking (IG: 22, rl: 21 y CH, SP: 17). Partly through a classifier, this position is kept with C4.5 and NB, but not with a NN in first position at 25% for Relief. From here through the whole features set, RL is the one that most frequently holds first position. At a 100% ranking, relief wins with a difference 25 times in comparison to CH, 12 to IG, and 10 to SP, and wins equally at 50% of ranking features. Results are kept with these percentages (50 and 100) for the three classifiers.

If we do the study regarding the entire eight tested by classifiers, there are no large differences. For C4.5 classifier, SP and RL methods stand out with very few differences regarding to IG. IG and RL are those that hold first positions with NB, while with 1NN it is SP. SOAP and Relief, with 152, are the ones which stayed in first position most of the time in all the tests (480); with IG 148, and with chi2 130, following.

We can adhere to the next recommendations due to the results obtained through the last three tests (AURC, AURC´s percentage and AURC with the first features of the arrange list): (I) AURC gives a more complete ranking goodness idea than the exactness obtained by a feature subset. (II) The complete best valued list is generated by the *RL* algorithm. However, if we are going to work with the first features, or with less than 25% of the features, SP and IG methods offer better results in less time. (III) In general, the best classification results are obtained by 1NN, although when the selected features number is smaller (less than the 25%), the performance of C4.5 was better in the four cases than in the rest of the classifiers.

6 Conclusions

Traditional work, where comparisons of feature ranking algorithms are made, mainly evaluate and compare the way of features selection instead of ranking methods. In this paper we present a methodology for evaluating ranking, beginning from the premise of no existence of any singular unique subgroup ideal for every case, and that the best ranking will depend on what the user is looking for.

We can conclude that the Area Under Ranking Classification Performance Curve ($AURC$) shows the complete performance of the orderly features list, globally indicating its predictive power. Based on the analysis of the evolution of AURC, we propose the use of algorithms SP and IG for C4.5 classifier with few features, and the use of RL with classifier 1NN in the rest of cases.

From here, our work aims to confirm if these results can be applied to other larger data sets as well as to study in depth if any relation exists between ranking method and selected classifier. Furthermore, we plan to increase our study with other measures of feature evaluation.

References

1. Liu, H., Motoda, H.: Feature Selection for Knowlegde Discovery and Data Mining. Kluwer Academic Publishers, London, UK (1998)
2. Hall, M., Holmes, G.: Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Eng. **15** (2003)
3. Ruiz, R., Riquelme, J., Aguilar-Ruiz, J.: Projection-based measure for efficient feature selection. Journal of Intelligent and Fuzzy System **12** (2002) 175–183
4. Hall, M.: Correlation-based feature selection for discrete and numeric class machine learning. In: 17th Int. Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA (2000) 359–366
5. Stoppiglia, H., Dreyfus, G., Dubois, R., Oussar, Y.: Ranking a random feature for variable and feature selection. Journal of Machine Learning Research **3** (2003) 1399–1414
6. Huang, J., Ling, C.: Using AUC and accuracy in evaluating learning algorithms. IEEE Transaction on Knowledge and data Engineering **17** (2005) 299–310
7. Yu, L., Liu, H.: Redundancy based feature selection for microarray data. In: 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. (2004)
8. Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric attributes. In: 7th IEEE Int. Conf. on Tools with Artificial Intelligence. (1995)
9. Kira, K., Rendell, L.: A practical approach to feature selection. In: 9th Int. Conf. on Machine Learning, Aberdeen, Scotland, Morgan Kaufmann (1992) 249–256
10. Kononenko, I.: Estimating attributes: Analysis and extensions of relief. In: European Conf. on Machine Learning, Vienna, Springer Verlag (1994) 171–182
11. Quinlan, J.R.: C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, California (1993)
12. Mitchell, T.: Machine Learning. McGraw Hill (1997)
13. Aha, D., Kibler, D., Albert, M.: Instance-based learning algorithms. Machine Learning **6** (1991) 37–66
14. Witten, I., Frank, E.: Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco (2000)
15. Blake, C., Merz, E.K.: UCI repository of machine learning databases (1998)