Feature Subset Selection for Learning Preferences: A Case Study

Antonio Bahamonde ANTONIO@AIC.UNIOVI.ES
Gustavo F. Bayón GBAYON@AIC.UNIOVI.ES
Jorge Díez JDIEZ@AIC.UNIOVI.ES
José Ramón Quevedo QUEVEDO@AIC.UNIOVI.ES
Oscar Luaces OLUACES@AIC.UNIOVI.ES
Juan José del Coz JUANJO@AIC.UNIOVI.ES
Jaime Alonso JALONSO@AIC.UNIOVI.ES
Artificial Intelligence Center, University of Oviedo at Gijón, E-33271 – Gijón (Asturias), Spain

Félix Goyache FGOYACHE@SERIDA.ORG
SERIDA-CENSYRA-Somió, C/ Camino de los Claveles 604, E-33203 Gijón (Asturias), Spain

Abstract

In this paper we tackle a real world problem, the search of a function to evaluate the merits of beef cattle as meat producers. The independent variables represent a set of live animals' measurements; while the outputs cannot be captured with a single number, since the available experts tend to assess each animal in a relative way, comparing animals with the other partners in the same batch. Therefore, this problem can not be solved by means of regression methods; our approach is to learn the preferences of the experts when they order small groups of animals. Thus, the problem can be reduced to a binary classification, and can be dealt with a Support Vector Machine (SVM) improved with the use of a feature subset selection (FSS) method. We develop a method based on Recursive Feature Elimination (RFE) that employs an adaptation of a metric based method devised for model selection (ADJ). Finally, we discuss the extension of the resulting method to more general settings, and provide a comparison with other possible alternatives.

1. Introduction

Learning preferences is a useful task in application fields like information retrieval, adaptive interfaces or quality assessment. The starting data set is a collection of preference judgments: pairs of vectors (v, u) where an agent expresses that it prefers v to u. In other words, training sets are samples of binary relations between objects described by the components of real number vectors.

This learning task can be accomplished following two approaches. We may look for classifiers to decide whether a pair (v, u) belongs or not to the relation, like in (Utgoff & Saxena, 1987; Branting & Broos, 1997). In general, the relation so induced is not transitive. However, Cohen et al. (1999) describe an algorithm that heuristically finds a good approximation to the ordering that best agrees with the learned binary relation.

The second approach tries to find an assessment or ranking function able to assign a real number to each vector in such a way that preferable objects obtain higher values. This point of view is followed in (Tesauro, 1988; Utgoff & Clouse, 1991; Herbrich et al., 1999; Fiechter & Rogers, 2000; Joachims, 2002; Díez et al., 2002); using different tools they propose algorithms to find a suitable assessment function, usually a linear function.

The main difficulty of the functional approach for learning preferences is that we do not have any class attached to training examples. So we can not use any regression method; instead, we can reduce the learning problem to separate two sets of vectors: positive vectors of the form $(v-u)$, and negative $-(v-u)$, for each preference judgment (v, u). Therefore, we can induce assessment functions using Support Vector Machines, SVM (Vapnik, 1998).

Appearing in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004. Copyright by the authors.

In this paper we present a real world assessment problem that motivates the development of tools for feature subset selection (FSS). The next section spells out the specific difficulties of finding an assessment for live beef cattle according to their merits as meet producers.

To face our FSS problem, we built tools that work in two stages. First, they produce an ordering or ranking of the features according to their usefulness. Here we discuss the use of RFE (Recursive Feature Elimination) (Guyon et al., 2002) comparing its achievements with those obtained by *Relieve*, a Kohavi and John (1997) modification of *Relief* (Kira & Rendell, 1992). The second stage is accomplished by a model selection method; it has to decide which subset of the k most useful features will produce higher accuracies. For this purpose, we consider a simple cross-validation (CV) estimation, and we introduce an adaptation of a metric based method (Schuurmans, 1997; Schuurmans & Southey, 2002) called *ADJusted distance estimate* (ADJ).

We will find that RFE outperforms Relieve in all tested circumstances. Additionally, in the beef cattle assessment problem, we obtained the best results when we use CV after RFE. However, our adaptation of ADJ, called Q_ADJ in Section 4.2, reaches only slightly worse scores. Moreover, Q_ADJ with RFE is the best method in a family of artificial data sets designed to test the utility of these methods in more general problems of learning preferences. Another important advantage of the method presented in this paper is that it is much faster than CV. This is a very important issue when a high number of features describe the objects that appear in the data sets of preference judgments.

2. The beef cattle assessment problem

The problem was to induce an assessment function for live beef cattle according to animals' carcass value. In fact, in animal breeding, conformation assessment is used as an indirect indicator of the animal's performance (Goyache et al., 2001b). So, the morphology of beef cattle is expected to be useful in evaluating the animals as meat producers.

Carcass conformation largely depends on live anatomy. However, this relationship is not direct because of the influence on shapes and volumes of the skin, subcutaneous fat and internal organs. Two major problems should be solved to find reliable rules relating animal dimensions and its ability to produce beef: accurate measurements of animals' bodies must be obtained, and animals must be assessed according to the estimation of their carcass values (Goyache et al., 2001a; Díez et al., 2003).

However, carrying out zoometry on an animal is a hard and risky task. The presence of humans disturbs animals increasing the error in the measurements. Therefore, to obtain accurate body measurements in a representative sample of animals we must perform an indirect zoometry by using digital images (Goyache et al., 2001b), see Figure 2. On the other hand, we must have a trained group of experts able to assess live animals' conformation following criteria used in bovine carcass markets. In this part we had the help of the experts of the Association of Breeders of Asturiana de los Valles (ASEAVA) . For a long time, our experts had been valuing animals of this beef cattle breed in a subjective manner. So, the strength and inertia of the traditional methods had to be overcome.

Our experts tend to grade their preferences in a relative way, comparing animals with the other partners in the same batch. So, there is a kind of *batch effect* that often biases their assessments. Thus, an animal surrounded by poor conformed cattle will probably obtain a higher score than if it were presented together with better bovines. From a computational point of view, this means that regression is not an acceptable method to induce an assessment function. Nevertheless, the knowledge of our experts can be reliably represented by means of orderings of small groups of animals according to experts' estimation of carcass values.

Using this methodology, we collected a set of 529 preference judgments of 128 different cows, and 395 preference judgments of 91 different bulls. Sexual dimorphism leads to different assessment criteria; so, data from bulls and cows have been considered as different training sets.

3. Learning linear preference assessments

Let us assume that

$$v_i > u_i : i = 1, \ldots, n$$

is a sample of an ordering relation in $\mathbb{R}^d$ called *preference* relation. Our aim is to find an ordering preserving (monotone) function $f : \mathbb{R}^d \to \mathbb{R}$ that will be called *assessment* or *ranking function*. In other terms, we look for an assessment f for d-dimension vectors such that maximize the probability of having $f(v) > f(u)$ whenever $v > u$.

Following (Herbrich et al., 1999; Fiechter & Rogers, 2000; Joachims, 2002; Díez et al., 2002), we define the assessment of a vector as its distance to an *assessment hyperplane* $\langle w, x \rangle = 0$. From a geometrical point of

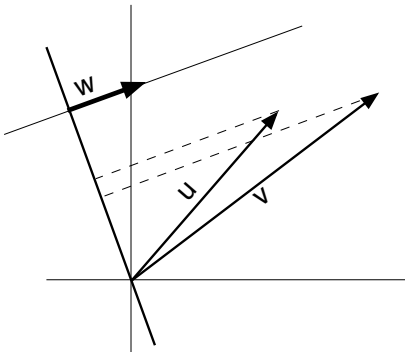

Figure 1. We are looking for a vector w such that the hyperplane $\langle w, x\rangle = 0$ is farther from preferable vectors. In the picture v is *better* than u, in symbols $v > u$

view, the function $f_w(x) = \langle w, x\rangle$ represents the distance to the hyperplane (of vectors perpendicular to w) multiplied by the norm of w, see Figure 1. The search of w is a NP-hard problem; however, it is possible to approximate the solution like in classification Support Vector Machines (Vapnik, 1998).

The core idea is that we can specify f_w taking into account that

$$f_w(v-u) > 0 \Leftrightarrow f_w(v) >_w f(u).$$

More formally, we have an optimization problem for margin maximization. We must minimize:

$$V(w,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

subject to:

$$\forall i = 1,\ldots,n,\ \langle w, v_i\rangle \geq \langle w, u_i\rangle + 1 - \xi_i$$

$$\forall i = 1,\ldots,n,\ \xi_i \geq 0$$

where C is a parameter that allows trading-off margin size against training error.

Additionally, as recommended in (Herbrich & Graepel, 2002), we will use SVM on normalized training examples. Therefore, the problem of finding a linear assessment function can be viewed as a problem of finding an hyperplane to separate the normalized differences: $\frac{v-u}{\|v-u\|}$ with class $+1$, and $-\frac{v-u}{\|v-u\|}$ with class -1, for each preference judgment $v > u$.

4. A FSS for learning preferences

In the preceding section, we showed that it is possible to induce an assessment or ranking function by means of an SVM that returns a hyperplane that separates two classes of vectors. Thus, we can use RFE (Guyon et al., 2002), a state-of-the-art algorithm, specially devised for SVM, that orders the set of features used to describe the training examples according to their usefulness to make an accurate classification rule. Then, we will use a model selection method to split the features list in order to obtain the most promising subset of features. For this purpose, we will introduce an adaptation of ADJ (Schuurmans & Southey, 2002), a metric-based method that has a natural implementation in our setting of learning preferences.

Algorithm 4.1 Pseudo code of SVM-RFE

Function SVM-RFE (T, fs) : A list of feature subsets
BEGIN
/* T: Set of training examples; each example is described by a vector of feature values (x) and its class (y)
fs: Set of features describing each example in T;
L: Ordered list of feature subsets; each subset contains the remaining features at every iteration; */
$F_d = fs$;
$L = [F_d]$; // Initially, one subset with all the features
for $j = d$ **downto** 2 **do**
  $\alpha = \text{SVM}(T)$; // Train SVM
  $w = \sum_k \alpha_k y_k x_k$; // w: the hyperplane coefficients
  $r = \arg\min_{i\in\{1,\ldots,|F_j|\}} ((w_i)^2)$; // The smallest ranking criterion
  $F_{j-1} = F_j \setminus f_r$; // Remove r-th feature from F_j
  $L = L + F_{j-1}$; // Add the subset of remaining features
  // Remove r-th feature from examples in T
  $T = \{x'_i : x'_i$ is $x_i \in T$ with f_r removed$\}$;
end for
// Return the ordered list L of feature subsets
return (L);
END

4.1. RFE in brief

RFE, which stands for Recursive Feature Elimination, is an example of a backward feature elimination process. So, it starts with all possible features and removes one feature per iteration, the one with the smallest *feature ranking criterion*, as shown in Algorithm 4.1. When the learner is a linear kernel SVM, RFE's criterion is the value of $(w_i)^2$, where w_i is the coefficient of the i-th feature in the separating hyperplane equation induced by SVM. A theoretical justification for using this criterion can be found in (Guyon et al., 2002).

This algorithm let us obtain a ranked list $L = (F_d, F_{d-1}, \ldots, F_1)$ with d different feature subsets, where each F_i is a subset with exactly i features. Due to the recursive elimination, features in a subset F_i are optimal in some sense when considered together, although individually they could be less relevant than

other features eliminated in a previous step. This is an interesting property of RFE since it takes into account possible relations between features, empowering the possibility of discovering useful groups of interrelated features that would be labeled as irrelevant if considered one by one. However, it should be noted that, given the greedy nature of RFE, F_i will not necessarily contain the i most useful features of the original feature set in order to achieve a higher accuracy.

4.2. The adaptation of ADJ

Once obtained the ranked list of feature subsets, the next step shall be to select one of them. In general, we will be interested in a subset which lets the learner yield the best performance, in terms of accuracy; so we need to estimate the performance for every feature subset.

This task can be accomplished by many different model selection techniques, for example, cross-validation (CV), a commonly used method that has been proved very reliable in many circumstances (Kohavi, 1995). However, CV is computationally costly. Moreover, it is also known that CV has high variance, which in some cases downgrades its performance as accuracy estimator. This disadvantage worsens as the number of training examples is reduced, what is frequent when we are learning preferences.

An alternative to CV and other accuracy estimators is a metric-based method called ADJ (Schuurmans, 1997; Schuurmans & Southey, 2002) devised to choose the appropriate level of complexity required to fit to data. In our case, given the nested sequence of feature sets provided by RFE, $F_1 \subset F_2 \subset \ldots \subset F_d$, ADJ would provide a procedure to select one of the hyperplanes g_i induced by SVM from the corresponding F_i.

The key idea is the definition of a metric on the space of hypothesis. Thus, for two hypothesis f and g, their distance is calculated as the expected disagreement in their predictions

$$d(f,g) \stackrel{\text{def}}{=} \varphi\left(\int err(f(x),g(x))d\mathrm{P}_X\right)$$

where $err(f(x),g(x))$ is the measure of disagreement on a generic point x in the space of example descriptions X. Given that these distances can only be approximated, ADJ establishes a method to compute $\hat{d}(g,t)$, an adjusted distance estimate between any hypothesis g and the *true* target classification function t. Therefore, the selected hypothesis is

$$g_k = \arg\min_{g_l} \hat{d}(g_l,t).$$

The estimation of distance, $\hat{d}$, is computed by means of the expected disagreement in the predictions in a couple of sets: the training set T, and a set U of unlabeled examples, that is, a set of cases sampled from P_X but for which the pretended *correct* output is not given. The ADJ estimation is given by

$$ADJ(g_l,t) \stackrel{\text{def}}{=} d_T(g_l,t)\cdot\max_{k<l}\frac{d_U(g_k,g_l)}{d_T(g_k,g_l)}$$

where, for a given subset of examples S, $d_S(f,g)$ is the expected disagreement of hypothesis f and g in S. Notice that we must avoid the impossibility of using the previous equation when there are zero disagreements in T for two hypotheses. Our proposal here is to use the Laplace correction to the probability estimation, in symbols

$$d_S(f,g) \stackrel{\text{def}}{=} \frac{1}{|S|+2}\left(1+\sum_{i\in S}1_{f(x_i)\neq g(x_i)}\right).$$

In general, it is not straightforward to obtain a set of unlabeled examples, so Bengio and Chapados (2003) proposed a sampling method over the available training set. However, for learning preferences, we can easily build the set of unlabeled examples from a set of preference judgments formed by pairs of real objects randomly selected from the original preference judgment pairs. We fix the size of U to be 10 times the size of T.

Our last modification of ADJ can only be used when we have more training examples than features; our data sets about beef cattle have this property, see Section 5.1 for more details. The idea of this proposal is borrowed from (Quinlan, 1992), and consists in adjusting the training errors, $d_T(g_l,t)$, taking into account the sizes of the linear problem, given that we are using linear surfaces to separate two classes. Thus, we introduce a ratio $Q = \frac{|T|+l}{|T|-l}$. Our intention is to penalize the scores achieved when the number of training examples, $|T|$, is near the number l of parameters in the model g_l. Finally,

$$\hat{d}(g_l,t) = Q\cdot ADJ(g_l,t) = \frac{|T|+l}{|T|-l}ADJ(g_l,t)$$

5. Experimental results

We conducted a set of experiments to show the benefits of our approach both in real world and artificial data sets. So, we established a comparison of the performance of two ranking methods endowed with two different procedures for selecting a feature subset. For

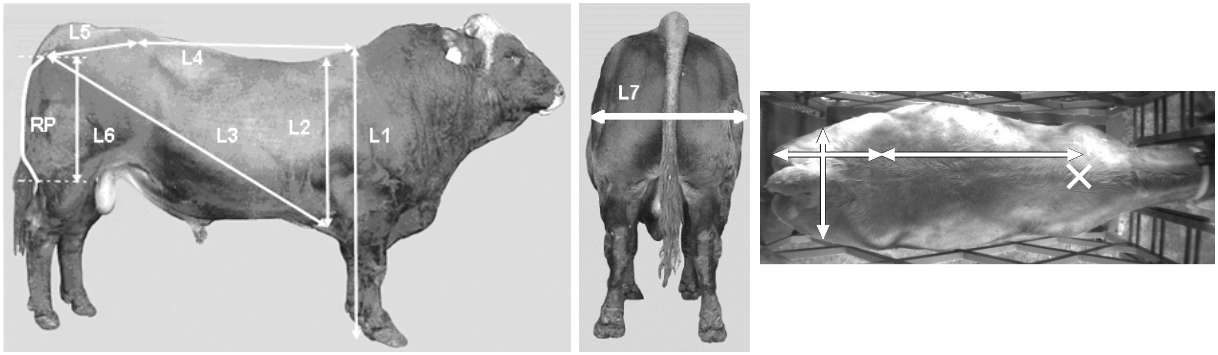

Figure 2. An example of indirect zoometry process using digital images. The leftmost two images, lateral and rear views, show 7 different lengths, plus the round profile (RP); from these features, a set of areas and volumes describe live animals' conformation. The right image is a zenithal view, one of the two stereo images that we are testing to use instead of the other views

the sake of completion we used a SVM to give a baseline measure of the accuracy that could be reached in each dataset.

In addition to RFE, we have implemented Relieve as a filter able to order the set of features that describe the examples of the dataset. To select the subset of the most useful features, we used ADJ and Q_ADJ as was explained in the previous section; as an alternative option we employed a classical cross validation performed in the training set. For implementing ADJ and Q_ADJ we used a set of unlabeled cases of size equal to 10 times the size of the training set. In all cases Q_ADJ outperformed ADJ in number of attributes; the scores in accuracy are similar in real word data sets, but Q_ADJ is significantly better in artificial data sets. We report the scores of ADJ and Q_ADJ in all cases; but, to ease the discussions that follow, we will only allude to Q_ADJ achievements.

In all cases we have used the $\text{SVM}^{\text{light}}$ implementation of Joachims (1999) with the default parameters and a linear kernel, but asking the system to find a separating hyperplane $\langle w, x\rangle + b = 0$ with $b = 0$. Additionally, the feature values in all training sets were normalized dividing their values by the typical deviation, notice that in our case the average of all features is zero.

Throughout all the experiments, all the inducers were run on identical training and test sets. On the other hand, we want to point out that the feature selection algorithms always used separated sets for training and testing, as was recalled in (Reunanen, 2003).

5.1. Real world data sets

The first package of experiments is taken from a real world application, presented in Section 2.

From each animal, we obtained 7 lengths from different parts of its body, plus the curvature of their *round profile* (RP), see Figure 2. To this set of 8 features, we added the sum of L5 and L4, since the whole measure of the top part could result useful independently of their component. In order to facilitate the acquisition of measurements, the length L3 was assumed to be the hypotenuse of a right-angled triangle formed by L4+L5 and L2. On the other hand, to try to describe faithfully the carcass merits, it is acknowledged that some volumes and areas can be very informative. Hence, adding all possible 2 and 3 dimensional data, each animal was described by 165 features. Additionally, we included all ratios in between the 8 lengths measured of each animal, resulting in other data sets with 193 features. We included these new 28 features since it is usually assumed that somehow harmonious proportions of body measurements are related to animals' performance. Nevertheless, our experimental results (see Tables 1 and 2) do not support significantly this idea.

Taking into account the complexity of obtaining the measurements from the lateral and rear views, we have considered the alternative of using only one stereo photograph from a zenithal point of view (see Figure 2), with the addition of the curvature of the RP. In this case, we do not have neither L6 nor L2; however, we observed that there is a high correlation between L3 and L4+L5, and then we can estimate L3 directly from L4+L5, and then we compute L2 using the right-angled triangle of these 3 lengths. Therefore, using this view we describe the animal by means of 7 lengths, and the curvature of the RP; finally, when we include the volumes and areas we have 120 features, and with the addition of ratios we have 141 features.

In all cases the methods that select the features to in-

Table 1. Classification accuracies estimated by a 10-fold cross-validation. We report here the scores achieved with Q_ADJ, and CV selections performed over a feature ordering obtained with Relieve. The column labeled with SVM represents the accuracies reached without any feature selection. Included in the names of data sets are the kind of animal (bull or cow), the view used to obtain the basic lengths (L for lateral or Z for zenithal), and the numbers of features used to describe live conformation of the animals

	Relieve + CV		Relieve + ADJ		Relieve + Q_ADJ		SVM
Dataset	%Acc.	#Feat.	%Acc.	#Feat.	%Acc.	#Feat.	%Acc.
BULLS-Z-120	95.43±2.76	9.30±5.37	94.42±3.20	10.50±10.62	94.42±3.20	5.90±3.99	94.17±2.79
BULLS-Z-141	95.44±2.97	12.40±5.94	94.42±1.94	13.20±9.83	94.67±2.15	8.20±5.51	94.68±2.89
BULLS-L-165	95.69±1.98	20.80±6.71	95.44±1.90	18.30±11.87	95.44±1.90	14.60±4.63	94.42±2.24
BULLS-L-193	96.45±2.04	25.40±11.24	95.69±2.57	25.20±9.89	95.69±2.57	22.10±8.22	94.68±2.41
COWS-Z-120	93.00±3.70	15.20±2.36	92.43±4.39	18.30±6.26	92.43±4.39	15.20±3.63	93.19±3.42
COWS-Z-141	93.19±3.43	16.30±8.74	92.80±4.60	20.70±17.58	92.80±4.60	12.20±6.66	92.81±3.60
COWS-L-165	93.19±3.72	42.60±27.63	93.56±3.63	51.10±54.12	93.37±3.53	18.20±3.63	93.00±3.30
COWS-L-193	93.37±3.22	23.30±11.32	93.56±3.10	21.00±20.77	92.81±2.81	9.40±1.91	93.00±3.30
Av.	94.47	20.66	94.04	22.29	93.95	13.23	93.74

duce an assessment function outperform the accuracy found by SVM, see Tables 1 and 2. Both in accuracy and number of features, the average differences between RFE+CV and all the other methods are statistically significant with $p < 0.05$ in 1-tail t tests. In the second position, RFE +Q_ADJ is significantly better than the rest with $p < 0.05$, except in the comparison with the accuracy of Relieve+CV where we only can assume that $p < 0.07$. Therefore, if we consider the differences between Relieve and RFE, they are clearly in favour of RFE, both in accuracy and in number of features selected. The comparison of Q_ADJ versus CV (with RFE) yields a slightly, although significant, higher accuracy in CV, while the difference in the number of features selected is more apparent than real given that in both cases the number of measurements required by the assessment functions is about 5.2, since the other features are areas, volumes or ratios.

The important question in practice about the feasibility of using one stereo zenithal view deserves a positive answer. The differences in accuracy and number of features are perfectly assumable.

5.2. Artificial data sets

We have made an in-depth study about the behaviour of RFE with CV and Q_ADJ in the presence of different levels of noise and number of relevant features. For this purpose we have designed a group of artificial data sets of 500 preference judgments where each object is described by 200 features with random values in the interval $[-1, +1]$. The name of each data set indicates the number of relevant features as well as the percentage of noise included. So, A-*R*-*N* refers to a problem with only R relevant attributes (varying from 10 to 40), and with a N% of noisy examples (from 0% to 20%). The assessment function used to order the preferences was $f(x) = \sum_{i=1}^{R} a_i x_i$; where a_i was randomly chosen as $+1$ or -1; in this way we ensure that only the first R features in each data set are equally relevant.

The scores of Q_ADJ (see Table 3) outperform those achieved by CV. The average differences in accuracy are statistically significant with $p < 0.04$ in a 1-tail t-test, in number of features the significance level is $p < 0.01$. Both methods improve significantly the results of SVM.

The experiments were repeated using data sets with different number of preference judgements, from 300 to 600, obtaining very similar results (omitted for lack of space) to those shown in Table 3.

6. Conclusions

In this paper we have dealt with preference judgments about objects whose descriptions need an important number of features. Our motivating case was to look for a function able to assess live beef cattle according to their carcass values. The conformation of each animal, the input of that function, can be considered as a vector whose components are profiles, lengths, areas, and volumes of different parts of their bodies. Due to the kind of knowledge available from the experts, this problem could not be solved by regression. Therefore, to discover an explicit formulation of this assessment

Table 2. Classification accuracies estimated by a 10-fold cross-validation. See caption of Table 1 for details

	RFE + CV		RFE + ADJ		RFE + Q_ADJ		SVM
Dataset	%Acc.	#Feat.	%Acc.	#Feat.	%Acc.	#Feat.	%Acc.
BULLS-Z-120	96.46±3.03	6.40±3.47	95.96±3.22	14.50±9.29	96.21±3.63	9.10±3.45	94.17±2.79
BULLS-Z-141	96.69±2.82	3.90±1.45	96.96±2.49	6.80±5.56	96.70±2.29	6.40±5.68	94.68±2.89
BULLS-L-165	96.20±3.45	4.50±1.28	95.70±2.99	24.10±25.51	95.44±3.56	6.60±2.97	94.42±2.24
BULLS-L-193	96.70±2.30	5.70±1.19	95.95±2.33	10.00±8.59	95.95±2.33	6.20±2.96	94.68±2.41
COWS-Z-120	94.14±2.60	4.90±1.45	93.57±3.50	4.20±1.17	93.57±3.50	4.20±1.17	93.19±3.42
COWS-Z-141	93.95±2.65	4.20±1.25	93.19±2.95	18.70±19.97	93.57±2.57	5.40±4.03	92.81±3.60
COWS-L-165	94.33±2.40	4.90±1.58	94.14±2.46	7.60±6.09	94.20±1.89	5.86±2.46	93.00±3.30
COWS-L-193	93.56±3.34	6.50±3.32	93.18±3.84	10.20±10.33	93.18±3.84	6.30±9.31	93.00±3.30
Av.	95.25	5.13	94.83	12.01	94.85	6.26	93.74

function, we learned a ranking map coherent with the *preferences* of the experts.

Thus, we collected 529 comparisons of cows, and 395 of bulls. Then, following (Herbrich et al., 1999; Fiechter & Rogers, 2000; Joachims, 2002; Díez et al., 2002), we reduced this problem to a binary classification that can be solved by means of a linear SVM. However, in order to improve both the accuracy and the descriptive power of the assessment, we designed some feature subset selection methods.

The best performance was achieved by those methods based on RFE (Guyon et al., 2002), that returns a sequence of models dealing with an increasing number of features. To decide the appropriate level of complexity required to fit to data, we have discussed the use of CV, and a new and much faster procedure called Q_ADJ, an adaptation of ADJ (Schuurmans, 1997; Schuurmans & Southey, 2002). Although in beef cattle, CV yields better results than Q_ADJ, the absolute differences are quite small. Additionally, we showed that this is not a general behavior; in fact, we provided a wide collection of data sets for learning preferences, where Q ADJ obtains significantly higher accuracy and less number of features. Therefore, we conclude that Q_ADJ is a reasonably alternative to CV.

Acknowledgements

The research reported in this paper has been supported in part under Spanish Ministerio de Ciencia y Tecnología (MCyT) and Feder grant TIC2001-3579. The authors would like to thank the Association of Breeders of Asturiana de los Valles (ASEAVA) for their help in the acquisition of beef cattle data sets.

References

Bengio, Y., & Chapados, N. (2003). Extensions to metric-based model selection. *Journal of Machine Learning Research*, *3*, 1209–1227.

Branting, K., & Broos, P. (1997). Automated acquisition of user preferences. *International Journal of Human-Computer Studies*, 55–77.

Cohen, W., Shapire, R., & Singer, Y. (1999). Learning to order things. *Journal of Artificial Intelligence Research*, *10*, 243–270.

Díez, J., Bahamonde, A., Alonso, J., López, S., del Coz, J., Quevedo, J., Ranilla, J., Luaces, O., Álvarez, I., Royo, L., & Goyache, F. (2003). Artificial intelligence techniques point out differences in classification performance between light and standard bovine carcasses. *Meat Science*, *64*, 249–258.

Díez, J., del Coz, J., Luaces, O., Goyache, F., Alonso, J., Peña, A., & Bahamonde, A. (2002). Learning to assess from pair-wise comparisons. *Procs. of the 8th IBERAMIA* (pp. 481–490). Sevilla, Spain.

Fiechter, C., & Rogers, S. (2000). Learning subjective functions with large margins. *Procs. of the 17th ICML* (pp. 287–294). Stanford, California, USA.

Goyache, F., Bahamonde, A., Alonso, J., López, S., del Coz J.J., Quevedo, J., Ranilla, J., Luaces, O., Alvarez, I., Royo, L., & Díez, J. (2001a). The usefulness of artificial intelligence techniques to assess subjective quality of products in the food industry. *Trends in Food Science & Technology*, *12*, 370–381.

Goyache, F., del Coz, J., Quevedo, J., López, S., Alonso, J., Ranilla, J., Luaces, O., Alvarez, I., & Bahamonde, A. (2001b). Using artificial intelligence to design and implement a morphological assessment system in beef cattle. *Animal Science*, *73*, 49–60.

Table 3. Results on artificial data sets with 500 examples and 200 features each one; their names A-*R*-*N* indicate the number of relevant features (R) and the percentage (N) of noisy examples

	RFE+CV		RFE+ADJ		RFE+Q_ADJ		SVM
Dataset	%Acc.	#Feat.	%Acc.	#Feat.	%Acc.	#Feat.	%Acc.
A-10-0	98.15	10	96.85	12	96.85	12	83.60
A-10-5	96.95	10	96.95	10	96.95	10	81.30
A-10-10	80.90	57	94.45	11	94.45	11	77.15
A-10-15	81.55	35	79	50	90.15	13	74.30
A-10-20	79.20	39	77.65	43	77.65	43	71.90
A-20-0	94.30	22	94.5	24	95.00	21	83.65
A-20-5	95.25	22	92.95	25	92.95	25	82.55
A-20-10	94.40	21	93.45	22	93.45	22	78.70
A-20-15	78.00	63	78.55	56	78.55	56	74.10
A-20-20	74.15	49	70.5	154	75.00	46	71.10
A-30-0	91.85	38	94.5	31	94.50	31	82.45
A-30-5	93.90	31	86.25	51	92.75	32	80.80
A-30-10	85.40	41	80.15	92	88.45	32	77.85
A-30-15	79.65	53	75.8	107	83.80	29	75.45
A-30-20	73.85	63	72.85	83	73.85	22	71.10
A-40-0	92.50	44	94.15	40	94.15	40	83.00
A-40-5	86.95	44	86.95	44	86.95	44	81.35
A-40-10	76.00	63	76.3	71	77.55	26	78.25
A-40-15	77.00	64	76.95	73	76.95	73	75.40
A-40-20	70.75	52	71.05	83	70.75	58	72.65
Av.	85.04	41.05	84.49	54.10	86.54	32.30	77.83

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*, 389–422.

Herbrich, R., & Graepel, T. (2002). A PAC-bayesian margin bound for linear classifiers. *IEEE Transactions on Information Theory*, 3140–3150.

Herbrich, R., Graepel, T., & Obermayer, K. (1999). Support vector learning for ordinal regression. *Procs. of the Ninth ICANN* (pp. 97–102). Edinburgh, UK.

Joachims, T. (1999). *Advances in kernel methods - support vector learning*, chapter Making Large-Scale SVM Learning Practical. MIT-Press.

Joachims, T. (2002). Optimizing search engines using clickthrough data. *Procs. of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. *Procs. of the Ninth ICML* (pp. 249–256).

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Procs. of the IJCAI* (pp. 1137–1145).

Kohavi, R., & John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*, 273–324.

Quinlan, J. (1992). Learning with continuous classes. *Proceedings 5th Australian Joint Conference on Artificial Intelligence* (pp. 343–348). Singapore.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, *3*, 1371–1382.

Schuurmans, D. (1997). A new metric-based approach to model selection. *AAAI/IAAI* (pp. 552–558).

Schuurmans, D., & Southey, F. (2002). Metric-based methods for adaptive model selection and regularization. *Machine Learning*, *48*, 51–84.

Tesauro, G. (1988). Connectionist learning of expert preferences by comparison training. *Advances in Neural Information Processing Systems (Proc. NIPS '88)* (pp. 99–106). MIT Press.

Utgoff, P., & Clouse, J. (1991). Two kinds of training information for evaluation function learning. *Procs. of the Ninth National Conference on Artificial Intelligence* (pp. 596–600). Anaheim, CA.

Utgoff, P., & Saxena, S. (1987). Learning a preference predicate. *Procs. of the 4^{th} International Workshop on Machine Learning* (pp. 115–121). Irvine, CA.

Vapnik, V. (1998). *Statistical learning theory*. John Wiley.