Analyzing sensory data using non-linear preference learning with feature subset selection[⋆]

Oscar Luaces, Gustavo F. Bayón, José R. Quevedo, Jorge Díez,
Juan José del Coz, and Antonio Bahamonde

Artificial Intelligence Center
University of Oviedo at Gijón
Campus de Viesques, s/n
E33271 – Gijón (Spain)
{oluaces,gbayon,quevedo,jdiez,juanjo,antonio}@aic.uniovi.es

Abstract The quality of food can be assessed from different points of view. In this paper, we deal with those aspects that can be appreciated through sensory impressions. When we are aiming to induce a function that maps object descriptions into ratings, we must consider that consumers' ratings are just a way to express their preferences about the products presented in the same testing session. Therefore, we postulate to learn from consumers' preference judgments instead of using an approach based on regression. This requires the use of special purpose kernels and feature subset selection methods. We illustrate the benefits of our approach in two families of real-world data bases.

1 Introduction

The quality of food can be assessed from different points of view. In this paper we are concerned with sensory quality from the perspective of consumers. This is a very important issue for food industries since they are aiming to adapt their production processes to improve the acceptability of their specialties. Thus, they need to discover the relationship between product descriptions and consumers' sensory degree of satisfaction. An excellent survey of the use of sensory data analysis in the food industry can be found in [1]; for a Machine Learning perspective, see [2,3].

From a conceptual point of view, sensory data can include the assessment of food products provided by two different kinds of groups of people usually called *panels*. The first one is made up of a small selected group of expert, trained judges; they will rate different aspects of products related to their taste, odor, color, etc... The most essential property of expert panelists, in addition to their discriminatory capacity, is their own coherence, not the uniformity of the group. We must assume that a given rating means the same for a given expert in every product; though not necessarily for every expert. Experts' panel will play the

[⋆] The research reported in this paper has been supported in part under Spanish Ministerio de Ciencia y Tecnología (MCyT) and Feder grant TIC2001-3579.

role of a bundle of sophisticated sensors; their ratings are used to describe each product, probably in addition to some chemical or physical devices.

The second kind of panel is made up of a group of untrained consumers (C); they are asked to rate their degree of acceptance or satisfaction about the tested products on a scale. Usually, this panel is organized in a set of *testing sessions*, where a group of potential consumers assess some instances from a sample E of the tested product. In general, each consumer only participates in a small number (sometimes only one) of testing sessions, usually in the same day.

In this paper we propose to tackle sensory data analysis by learning *consumers' preferences*, see [4,5,6] where training examples will be represented by *preference judgments*: pairs of vectors $(\boldsymbol{v}, \boldsymbol{u})$ where someone expresses that prefers the object represented by $\boldsymbol{v}$ to the object represented by $\boldsymbol{u}$. We will show that this approach can induce more useful knowledge than other approaches, like regression based methods. The main reason is due to the fact that preference judgments sets can represent more relevant information to discover consumers' preferences.

At the end of the paper we show experimental results of preference learning in two real-world data bases taken from sensory data analysis of beef meat and traditional Asturian cider. In both cases, non-linear preference functions can explain consumers' preferences better than other methods. Additionally, as happens with any other machine learning application, feature subset selection (FSS) plays a very important role. In fact, sometimes FSS marks the difference between useful tools and merely academic developments [3]. In this paper we show how it is possible to adapt to preference learning some state of the art FSS methods designed for SVM (Support Vector Machines) [7] with non-linear kernels.

2 Why using preference learning?

Initially, sensory data can be viewed as in regression problems: the sensory descriptions (human and mechanical) of each object $\boldsymbol{x} \in E$ are endowed with a rating $r(\boldsymbol{x})$ that represents the degree of satisfaction for each consumer or the average value for a group of them. So, a straightforward approach to handle sensory data can be based on regression. However, this is not a faithful capturing of people's preferences [8,9]. The main reason is due to the fact that sensory data, expressed as a regression problem, do not represent all available knowledge. In particular, we would like to remark that consumers' ratings are just a way for expressing a relative ordering. There is a kind of *batch effect* that often biases the ratings so that a product will obtain a higher/lower rating when it is assessed together with other products that are clearly worse/better. Therefore, we must consider as a very important issue the information about the batches tested by consumers in each rating session.

Traditionally the process given to these data sets includes testing some statistical hypothesis [10,1]. On the other hand, the approach followed in [2] is based on the use of Bayesian belief networks. In both cases, these approaches demand

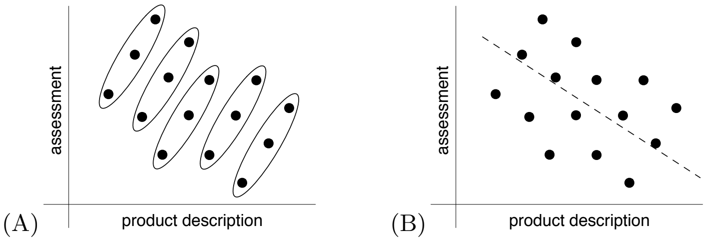

Figure 1. Each ellipse in (A) represents the assessments for a given session, where the assessment function is clearly different than the one obtained in (B) by a regression method applied to the whole set of assessments without information about sessions

that all available food products (the objects x) must be rated by all consumers; in practice, this is an impossible assumption most of the times. In general, each consumer will only assess a small number of products. Thus, we will have sets of ratings $(r_i(\boldsymbol{x}) : \boldsymbol{x} \in E_i)$ for each consumer or group of consumers $i \in C$, where $\cup(E_i : i \in C) = E$. In addition to this fact, let us emphasize some important peculiarities of the whole data collected in a sensory study that we have to take into account: i) we have different scales in the ratings, given that the assessments come from different sets of consumers; additionally, ii) these ratings suffer the batch effect alluded to previously.

The importance of these factors is graphically depicted in Figure 1. Here there is a collection of consumers' assessments (represented in the vertical axis) about some products whose descriptions are given by a single number x represented in the horizontal axis. If we observe Figure 1A, where the assessments of the same session are drawn inside ellipses, we can say that in each session the message of the consumers is the same: the more x the better. However, there are discrepancies about how this knowledge is expressed in different sessions. Probably because there are different consumers in each session; or perhaps because the same consumer forgets the exact number used to assess a given degree of satisfaction; or the sensory reactions were forgotten from one session to another.

If we do not consider sessions, the data collected become the cloud of points represented in Figure 1B. Then, it will be difficult for a regression method to discover the unanimous opinion of consumers. In fact, in this case, regression methods will conclude that the more x the worse, since that seems to be the general orientation of those points in the space. Therefore, the information about the sessions must be integrated in the data to be processed with the rest of sensory opinions and descriptions of the products tested by consumers. In the next section we will present our approach to deal with sessions explicitly. The overall idea is to avoid trying to predict the exact value of consumer ratings; instead we will look for a function that returns higher values to those products with higher ratings.

3 Learning preferences: an SVM approach

Although there are other approaches to learn preferences, following [4,5,6] we will try to induce a real *preference* or *ranking function* f from the space of objects considered, say $\mathbb{R}^d$, in such a way that it maximizes the probability of having $f(\boldsymbol{v}) > f(\boldsymbol{u})$ whenever $\boldsymbol{v}$ is preferable to $\boldsymbol{u}$. This functional approach can start from a set of objects endowed with a (usually ordinal) rating, as in regression, but essentially, we need a collection of preference judgments

$$PJ = \{\boldsymbol{v}_j > \boldsymbol{u}_j : j = 1, \ldots, n\} \tag{1}$$

When we have a family of ratings $(r_i(\boldsymbol{x}) : \boldsymbol{x} \in E_i)$ for $i \in C$, we transform them into a preference judgments set PJ considering all pairs $(\boldsymbol{v}, \boldsymbol{u})$ such that objects $\boldsymbol{v}$ and $\boldsymbol{u}$ were presented in the same session to a given consumer i, and $r_i(\boldsymbol{v}) > r_i(\boldsymbol{u})$. Hence, without any lost of generality, we can assume a set PJ as in formula (1).

In order to induce the ranking function, we can use the approach presented by Herbrich et al. in [4]. So, we look for a function $F : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d, F(\boldsymbol{x}, \boldsymbol{y}) > 0 \Leftrightarrow F(\boldsymbol{x}, \boldsymbol{0}) > F(\boldsymbol{y}, \boldsymbol{0}) \tag{2}$$

Then, the ranking function $f : \mathbb{R}^d \to \mathbb{R}$ can be simply defined by

$$\forall \boldsymbol{x} \in \mathbb{R}^d, f(\boldsymbol{x}) = F(\boldsymbol{x}, 0) \tag{3}$$

Given the set of preference judgments PJ (1), we can specify F by means of the constraints

$$F(\boldsymbol{v}_j, \boldsymbol{u}_j) > 0 \text{ and } F(\boldsymbol{u}_j, \boldsymbol{v}_j) < 0, \quad \forall j = 1, \ldots, n \tag{4}$$

Therefore, we have a binary classification problem that can be solved using SVM. If we represent preference judgments pairs $(\boldsymbol{v}, \boldsymbol{u})$ in a higher dimensional feature space by means of $(\phi(\boldsymbol{v}), \phi(\boldsymbol{u}))$, we will obtain a function of the form:

$$F(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} \alpha_i z_i \mathcal{K}(\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}, \boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{w}, (\phi(\boldsymbol{x}), \phi(\boldsymbol{y})) \rangle \tag{5}$$

where the pairs $\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}$ are the support vectors; $\boldsymbol{w}$ is the vector of weights in the higher dimensional feature space; and $\mathcal{K}$ is the kernel used by SVM. The key idea of this approach is the definition of the kernel $\mathcal{K}$ as follows

$$\mathcal{K}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4) = k(\boldsymbol{x}_1, \boldsymbol{x}_3) - k(\boldsymbol{x}_1, \boldsymbol{x}_4) - k(\boldsymbol{x}_2, \boldsymbol{x}_3) + k(\boldsymbol{x}_2, \boldsymbol{x}_4) \tag{6}$$

where k is a kernel function defined as the inner product of two objects represented in the feature space, that is, $k(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle$. In this case, it is easy to proof that F fulfills the conditions expressed in equation (2). In the experiments reported in Section 5, we will employ a polynomial kernel, defining

$k(\boldsymbol{x}, \boldsymbol{y}) = (\langle \boldsymbol{x}, \boldsymbol{y} \rangle + c)^g$, with $c = 1$ and $g = 2$. Notice that, in general, according to the previous definitions,

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i z_i (k(\boldsymbol{x}_i^{(1)}, \boldsymbol{x}) - k(\boldsymbol{x}_i^{(2)}, \boldsymbol{x})) \qquad (7)$$

Hence, for the polynomial kernel we will obtain a non-linear function that assesses the ranking for each object $\boldsymbol{x}$.

4 FSS in non-linear preference learning

A major issue when dealing with real-world problems involving sensory data is to find out those features which have more influence on the tastes of consumers; thus, the production process can focus on them to improve the acceptability of the final product. Additionally, reducing the number of features describing objects decreases the cost of data acquisition, which in many cases can make these machine learning techniques applicable in industrial processes [3].

In recent years several methods related to feature selection when using SVM have been developed. One of the most remarkable is RFE (Recursive Feature Elimination) [11]. Given a data set with objects described by a set of d features, $\mathcal{F}_d$, the method considers that i is the less useful feature if $|\boldsymbol{w}_i|$ is the smallest weight (see eq. 5). Then this feature is removed, giving rise to a subset $\mathcal{F}_{d-1}$ with $d-1$ features. The process is successively repeated until no more features are left. Notice that in this way, we obtain a ranking of the original d features, and a sequence of models, each one obtained using the corresponding subset of i features, $\mathcal{F}_i$. A chunk of features can also be removed instead of only one at each iteration, as suggested in [11].

However, RFE's criterion is not directly applicable for non-linear kernels, so we have used two state of the art methods to achieve ordered lists of features in non-linear scenarios. Moreover, we must take into account an important peculiarity of preference learning data sets. In this case, the examples are formed by pairs of objects $(\boldsymbol{v}, \boldsymbol{u})$, and both objects are described by the same set of d features. Therefore, somehow we must consider twice the merits of each feature to be removed and, in each iteration, we have to get rid of the two copies of the selected feature.

4.1 Feature ranking methods for non-linear preference kernels

Method 1.- The first method that we have applied to obtain a ranking of features with non-linear kernels was proposed by Rakotomamonjy [12]; its ranking criterion orders the list of features according to their influence in the variations of feature's weight; in fact, it is an extension of RFE to the non-linear case. In symbols, the method removes in each iteration the feature with the lowest ranking value:

$$R_1(i) = |\nabla_i \|\boldsymbol{w}\|^2| = \left| \sum_{k,j} \alpha_k \alpha_j z_k z_j \frac{\partial K(\boldsymbol{s}\cdot\boldsymbol{x}_k, \boldsymbol{s}\cdot\boldsymbol{x}_j)}{\partial s_i} \right|, \quad i = 1, \ldots, d \qquad (8)$$

where $\boldsymbol{s}$ is a scaling factor used to simplify the computation of partial derivatives. Given that we are facing a preference learning problem, where every example is a preference judgment like in (1), then we must modify the use of $\boldsymbol{s}$: we need 4 copies, one for each object involved in the definition of the kernel. Thus, according to (6), we compute

$$\begin{aligned}\frac{\partial \mathcal{K}(\boldsymbol{s}\cdot\boldsymbol{x}_1, \boldsymbol{s}\cdot\boldsymbol{x}_2, \boldsymbol{s}\cdot\boldsymbol{x}_3, \boldsymbol{s}\cdot\boldsymbol{x}_4)}{\partial s_i} &= \frac{\partial k(\boldsymbol{s}\cdot\boldsymbol{x}_1, \boldsymbol{s}\cdot\boldsymbol{x}_3)}{\partial s_i} - \\ -\frac{\partial k(\boldsymbol{s}\cdot\boldsymbol{x}_1, \boldsymbol{s}\cdot\boldsymbol{x}_4)}{\partial s_i} &- \frac{\partial k(\boldsymbol{s}\cdot\boldsymbol{x}_2, \boldsymbol{s}\cdot\boldsymbol{x}_3)}{\partial s_i} + \frac{\partial k(\boldsymbol{s}\cdot\boldsymbol{x}_2, \boldsymbol{s}\cdot\boldsymbol{x}_4)}{\partial s_i} \qquad (9)\end{aligned}$$

In this formula, for a polynomial kernel $k(\boldsymbol{x}, \boldsymbol{y}) = (\langle \boldsymbol{x}, \boldsymbol{y} \rangle + c)^g$ and a vector $\boldsymbol{s}$ such that $\forall i, s_i = 1$ we have that

$$\frac{\partial k(\boldsymbol{s}\cdot\boldsymbol{x}, \boldsymbol{s}\cdot\boldsymbol{y})}{\partial s_i} = 2g(x_i y_i)(c + \langle \boldsymbol{x}, \boldsymbol{y} \rangle)^{g-1} \qquad (10)$$

Method 2.- The second method was developed by Degroeve et al. [13] for splice site prediction of DNA sequences. This method uses a ranking criterion such that features are ordered with respect to the loss in predictive performance when they are removed. In [13] the authors approximate the generalization performance when removing the i-th feature by the accuracy on the training set while setting the value of that feature, in every instance, to its mean value. When using this method for preference learning with the kernel of equation (6) the ranking criterion can be expressed as

$$R_2(i) = \left(\sum_k z_k \cdot \sum_j \alpha_j z_j \mathcal{K}(\boldsymbol{x}_j^{(1),i}, \boldsymbol{x}_j^{(2),i}, \boldsymbol{x}_k^{(1),i}, \boldsymbol{x}_k^{(2),i} \right) \qquad (11)$$

where $\boldsymbol{x}^i$ denotes a vector describing an object where the value for the i-th feature was replaced by its mean value. Notice that a higher value of $R_2(i)$, that is, a higher accuracy on the training set when removing feature i-th, means a lower relevance of that feature. Therefore, we will remove the feature yielding the highest ranking value, as opposite to the ranking method described previously.

4.2 Model selection

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 $\mathcal{F}_i$ which lets the learner yield the best performance, in terms of accuracy; so we need to estimate the performance for every feature subset.

Following the same approach as in [6], we will not use cross-validation for this purpose; its main disadvantages are its computational cost as well as its high variance, so we will use an alternative model selection: ADJ, a metric-based method [14] 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 any of the ranking methods described previously, $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}_d$, ADJ would provide a procedure to select one of the models f_i induced by SVM from the corresponding $\mathcal{F}_i$.

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

$$d(f,g) \overset{\text{def}}{=} \varphi\left(\int err(f(\boldsymbol{x}), g(\boldsymbol{x})) d\mathrm{P}_X\right) \tag{12}$$

where $err(f(\boldsymbol{x}), g(\boldsymbol{x}))$ is the measure of disagreement on a generic point $\boldsymbol{x}$ in the input space 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 f and the *true* target classification function t. Therefore, the selected hypothesis is

$$f_k = \arg\min_{f_l} \hat{d}(f_l, t) \tag{13}$$

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

$$ADJ(f_l, t) \overset{\text{def}}{=} d_T(f_l, t) \cdot \max_{k<l} \frac{d_U(f_k, f_l)}{d_T(f_k, f_l)} \tag{14}$$

where, for a given subset of examples S, $d_S(f,g)$ is the expected disagreement of hypothesis f and g in S. To avoid the impossibility of using the previous equation when there are zero disagreements in T for two hypotheses we propose to use the Laplace correction to the probability estimation; thus,

$$d_S(f,g) \overset{\text{def}}{=} \frac{1}{|S|+2}\left(1 + \sum_{x \in S} 1_{f(\boldsymbol{x}) \neq g(\boldsymbol{x})}\right) \tag{15}$$

In general, it is not straightforward to obtain a set of unlabeled examples, so [15] proposed a sampling method over the available training set. However, for learning preferences, we can easily build the set of unlabeled examples: new preference judgment pairs can be formed by arranging real objects randomly selected from the original data.

4.3 Dealing with redundant features

As we have previously pointed out, sensory data include ratings of experts for different characteristics of the assessed products; it is not rare that several experts have similar opinions about a given characteristic. Some physical and chemical features can also present this kind of similarities. Therefore, these data sets may frequently present a certain degree of redundancy to describe an object more precisely. Trying to take advantage of these redundancies, we have developed a simple but quite effective filtering process, `RF`, to be applied to sensory data sets before any other feature subset selection process. On the other hand, this filter provides an additional benefit for feature selection algorithms, since the number of features to deal with is reduced.

`RF` is an iterative process where in each step the two most *similar* features are replaced by a new one whose values are computed as the average of them. Considering two features represented by $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ as (column) vectors whose dimension is the number of examples in the data set, the similarity can be estimated by means of their cosine; that is,

$$\text{similarity}(\boldsymbol{a}_i, \boldsymbol{a}_j) = \frac{\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle}{\|\boldsymbol{a}_i\| \cdot \|\boldsymbol{a}_j\|} \qquad (16)$$

Applying this method, we obtain a sequence of different descriptions of the original data set, each one with one feature less than the previous. To select an adequate description in terms of prediction accuracy, we use again ADJ. The selected description can be considered a summarized version of the original data set to be processed by the feature subset selection methods previously described.

5 Experimental results

To illustrate the benefits of our approach, we have conducted some experiments with a couple of sensory data bases. The first one comes from a study carried out to determine the features that entail consumer acceptance of beef meat from seven Spanish breeds [16]. Each piece of meat was described by: weight of the animal, aging time, breed, 6 physical features describing its texture and 12 sensory characteristics rated by 11 different experts (132 ratings). Given that breed was represented by 7 boolean features, the whole description of each piece of meat uses 147 features. In each testing session, 4 or 5 pieces of meat were tested and a group of consumers were asked to rate only three different qualities: tenderness, flavor and acceptance. These three data sets have over 2420 preference judgments.

The second data base deals with sensory data about traditional Asturian cider [17]. In this case, the description of each cider was given just by 64 chemical and physical features, without any expert rating. In fact, the consumers here were a set of 14 candidates to become experts, and the rating sessions (of 3, 4 or 5 ciders) were taken during the training and selection stage. These potential experts were asked to rate a high number of qualities of ciders: bouquet, color, acidity, bitterness, 4 additional visual aspects and 3 more flavor related aspects. Thus, we have 12 qualities of cider, that is, 12 different data sets of over 225 preference judgments.

5.1 Preference learning vs. regression

First, we performed a comparison between the scores achieved by preference approaches and those obtained by regression methods. As was explained in Section 2, the core point of preference learning approach is the concept of testing session. Thus, for each session, to summarize the opinions of consumers, we computed the mean of the ratings obtained by each food product, which was endowed to the objects' descriptions to conform the regression training sets. These sets can be used to induce a function that predicts numerical ratings of consumers. We have experimented with a simple linear regression and with a well reputed regression algorithm: Cubist, a commercial product from RuleQuest Research.

To interpret regression results we used the relative mean absolute deviation (*rmad*), which is computed from the mean absolute distance or deviation, *mad*, of the function f learned by the regression method:

$$\mathrm{rmad}(f) = 100 \cdot \frac{\mathrm{mad}(f)}{\mathrm{mad}(mean)} \tag{17}$$

where *mean* is the constant predictor that returns the mean value in all cases.

On the other hand, we can obtain some preference judgments from the ratings of the sessions comparing the rating of each product with the rest, one by one, and constructing the corresponding pair. To learn from preference judgment data sets, we used $\mathrm{SVM}^{\mathrm{light}}$ [18] with linear and polynomial kernels. In this case, the errors have a straightforward meaning as misclassifications; so in order to allow a fair comparison between regression and preference learning approaches, we also tested regression models on preference judgments test sets, calculating their misclassifications.

Table 1 reports the 10-fold cross validation scores achieved with the real-world data sets described, both with regression and preference learning methods. The scores show that regression methods are unable to learn any useful knowledge: their relative mean absolute deviation (*rmad*) is above 100% in almost all cases, that is, usually the mean predictor performs better. Even when these regression models are tested on preference judgment sets, the percentage of misclassifications is over 40%, clearly higher than those obtained when using the preference learning approach. SVM-based methods can reduce these errors up to an average near 30% with a linear kernel ($\mathtt{SVM}_l$ with $k(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$), and near 20% if the kernel is a polynomial of degree 2 ($\mathtt{SVM}_p$ with $k(\boldsymbol{x}, \boldsymbol{y}) = (\langle \boldsymbol{x}, \boldsymbol{y} \rangle + 1)^2$). The rationale behind the improvement, when using non-linear kernels, can be explained taking into account that the positive appreciation of food products usually requires an equilibrium of its components, and the increase or decrease of any value from that point is frequently rejected.

5.2 FSS in non-linear preference learning

In order to find out those features which have more influence on the tastes of consumers, we have applied the feature subset selection methods described in Section 4. For the sake of simplicity, in what follows $\mathtt{FSS}_1$ and $\mathtt{FSS}_2$ will denote

Table 1. Results on cider and beef meat data sets. For regression methods we report the relative mean absolute deviation; for preference learning, the percentage of preference judgments pairs misclassified is shown. The number of selected features is also shown for the `FSS` algorithms. Let us recall that original cider and meat data sets have 64 and 147 features, respectively. All these results have been obtained by a 10-fold cross-validation.

	Regression		Preferences				Preferences ($\texttt{SVM}_p$+`FSS`)									
	Linear	Cubist	Linear	Cubist	$\texttt{SVM}_l$	$\texttt{SVM}_p$	$\texttt{FSS}_1$		$\texttt{FSS}_2$		`RF`		$\texttt{RF+FSS}_1$		$\texttt{RF+FSS}_2$	
	Rmad	Rmad	Error	Error	Error	Error	Error	#Att.	Error	#Att.	Error	#Att.	Error	#Att.	Error	#Att.
acidity	103.0%	109.4%	40.0%	42.4%	29.9%	18.0%	17.2%	18.2	17.6%	16.9	16.4%	47.5	14.7%	19.3	19.3%	18.5
bitterness	105.8%	111.9%	56.0%	47.4%	30.5%	23.1%	25.1%	29.4	21.2%	22.0	20.8%	39.0	18.2%	23.9	22.9%	27.6
flavor-1	105.3%	111.7%	42.4%	44.3%	27.2%	17.1%	18.5%	30.0	21.4%	26.2	20.6%	36.0	20.1%	24.1	19.7%	22.2
flavor-2	107.2%	116.0%	45.6%	45.0%	28.6%	17.9%	19.1%	27.0	15.6%	26.3	17.8%	40.5	17.8%	22.9	16.5%	18.1
flavor-3	110.3%	107.7%	43.8%	41.8%	33.6%	17.7%	22.7%	28.8	21.8%	19.8	19.3%	30.0	20.1%	17.6	18.4%	18.3
bouquet	104.0%	110.2%	43.5%	42.7%	26.4%	21.0%	16.7%	30.0	18.9%	28.1	19.8%	45.0	18.8%	24.1	18.0%	23.1
color	98.4%	109.9%	41.3%	43.4%	26.1%	17.8%	19.5%	32.0	22.0%	22.6	19.5%	41.5	21.6%	24.3	24.9%	21.0
visual-1	103.2%	113.0%	41.7%	43.1%	25.9%	13.4%	11.5%	30.4	13.8%	24.5	12.0%	34.5	14.7%	24.5	13.8%	18.1
visual-2	102.3%	112.0%	43.8%	45.7%	34.0%	20.0%	21.1%	30.5	18.9%	23.8	19.9%	35.5	21.2%	27.1	19.4%	23.2
visual-3	107.2%	120.5%	45.6%	49.3%	25.3%	20.6%	16.1%	18.4	15.6%	18.0	13.8%	32.5	13.4%	24.6	13.4%	25.3
visual-4	98.7%	97.2%	36.5%	38.2%	23.0%	14.0%	14.1%	25.5	15.0%	19.4	14.1%	38.5	14.7%	22.4	12.1%	19.2
Average cider	104.1%	110.9%	43.7%	43.9%	28.2%	18.2%	18.3%	27.3	18.3%	22.5	17.6%	38.2	17.7%	23.2	18.0%	21.3
tenderness	96.3%	97.8%	41.5%	43.1%	29.6%	19.4%	-	-	-	-	20.0%	50.0	21.8%	27.0	21.3%	37.5
flavor	99.3%	103.4%	43.8%	46.5%	32.7%	23.8%	-	-	-	-	25.0%	65.0	26.5%	33.5	26.1%	29.0
acceptance	94.0%	97.2%	38.4%	40.2%	31.9%	22.1%	-	-	-	-	24.7%	39.5	24.8%	30.0	25.3%	26.7
Average meat	96.5%	99.5%	41.2%	43.3%	31.4%	21.8%	-	-	-	-	23.2%	51.5	24.4%	30.2	24.2%	31.1
Total average	102.5%	108.4%	43.1%	43.8%	28.9%	19.0%	-	-	-	-	18.8%	41.1	19.2%	24.7	19.4%	23.4

the selectors that use ranking *Method 1* and *Method 2* respectively. Additionally, we used RF in two senses: as a feature subset selector, and as a filter to be applied before FSS$_1$ and FSS$_2$. In all cases we used ADJ to choose among the subsets of features. The learner used was SVM$_p$, given that it was the most accurate in our tests. On the beef meat data sets it is almost impractical to use FSS$_1$ and FSS$_2$ due to its computational cost, unless a previous reduction in the number of features can be achieved; therefore we only have results for this data sets when RF is used as a previous filter. Moreover, features were processed in chunks of five for the meat data sets, while they were removed one by one for the cider data sets.

We can see (Table 1) that FSS$_1$, FSS$_2$, and RF considerably reduce the number of features without (in general) loss of accuracy. In the cider data sets, all methods obtain similar accuracy scores (non-significant differences), but FSS$_2$ is significantly better than FSS$_1$ reducing the number of features, while RF achieves the poorest scores in this task. For the cider data sets, accuracy scores obtained by FSS$_1$ and FSS$_2$ are slightly improved when RF is previously used. However, for the meat data sets, accuracy decreases slightly when we use the RF filter with respect to the accuracy obtained on the original data set by SVM$_p$; it also decreases when using FSS$_1$ and FSS$_2$ after RF. We think that this behavior is due to the the fact that we are removing chunks of five features in each iteration.

6 Conclusions

The analysis of sensory data is a very useful tool for food industries because it provides the knowledge to satisfy the tastes of consumers. These data sets present some peculiarities that make difficult the use of regression based algorithms: each consumer does not rate all available products; and they give numerical assessments only as a way to express a relative preference in a rating or testing session (batch effect).

Preference learning does not try to learn the exact rating; however, it finds out models able to explain consumer preferences. We have observed that the accuracy increases significantly with non-linear functional models in the two real-world data bases analyzed. In general, the usefulness of these models can be improved with the use of specially fitted FSS methods.

Another interesting peculiarity of sensory data sets is that, frequently, there are blocks of features describing the same aspect. To take advantage of these redundancies we have developed a filtering process that can be applied to improve the performance of the learner.

Acknowledgements

We would like to thank Carlos Sañudo (Animal Production at the University of Zaragoza) for providing us with the beef meat data sets [16]; we also thank Anna Picinelli (Cider group at SERIDA, Asturias) for allowing us to use cider data sets [17].

References

1. Murray, J., Delahunty, C., Baxter, I.: Descriptive sensory analysis: past, present and future. Food Research International **36** (2001) 461–471
2. Corney, D.: Designing food with bayesian belief networks. In: Proceedings of the International Conference on Adaptive Computing in engineering Design and Manufacture. (2002) 83–94
3. Goyache, F., Bahamonde, A., Alonso, J., López, S., del Coz J.J., Quevedo, J., Ranilla, J., Luaces, O., Alvarez, I., Royo, L., Díez, J.: The usefulness of artificial intelligence techniques to assess subjective quality of products in the food industry. Trends in Food Science and Technology **12** (2001) 370–381
4. Herbrich, R., Graepel, T., Obermayer, K.: Support vector learning for ordinal regression. In: Proceedings of the Ninth International Conference on Artificial Neural Networks, Edinburgh, UK (1999) 97–102
5. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD). (2002)
6. Bahamonde, A., Bayón, G.F., Díez, J., Quevedo, J.R., Luaces, O., del Coz, J.J., Alonso, J., Goyache, F.: Feature subset selection for learning preferences: A case study. In: Proceedings of the International Conference on Machine Learning (ICML '04), Banff, Alberta (Canada). (2004)
7. Vapnik, V.: Statistical Learning Theory. John Wiley, New York, NY (1998)
8. Cohen, W., Shapire, R., Singer, Y.: Learning to order things. Journal of Artificial Intelligence Research **10** (1999) 243–270
9. Dumais, S., Bharat, K., Joachims, T., Weigend, A., eds.: Workshop on implicit measures of user interests and preferences. In ACM SIGIR Conference, Toronto, Canada (2003)
10. Næs, T., Risvik, E.: Multivariate analysis of data in sensory science. Elsevier (1996)
11. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46** (2002) 389–422
12. Rakotomamonjy, A.: Variable selection using SVM-based criteria. Journal of Machine Learning Research **3** (2003) 1357–1370
13. Degroeve, S., De Baets, B., Van de Peer, Y., Rouzé, P.: Feature subset selection for splice site prediction. Bioinformatics **18** (2002) 75–83
14. Schuurmans, D., Southey, F.: Metric-based methods for adaptive model selection and regularization. Machine Learning **48** (2002) 51–84
15. Bengio, Y., Chapados, N.: Extensions to metric-based model selection. Journal of Machine Learning Research **3** (2003) 1209–1227
16. Gil, M., Serra, X., Gispert, M., Oliver, M., Sañudo, C., Panea, B., Olleta, J., Campo, M., Oliván, M., Osoro, K., Garcia-Cachan, M., Cruz-Sagredo, R., Izquierdo, M., Espejo, M., Martín, M., Piedrafita, J.: The effect of breed-production systems on the myosin heavy chain 1, the biochemical characteristics and the colour variables of longissimus thoracis from seven spanish beef cattle breeds. Meat Science **58** (2001) 181–188
17. Picinelli, A., Suárez, B., Moreno, J., Rodríguez, R., Caso-García, L., Mangas, J.: Chemical characterization of Asturian cider. Journal of Agricultural and Food Chemistry **48** (2000) 3997–4002
18. Joachims, T.: Making large-scale support vector machines learning practical. In B. Schölkopf, C. Burges, A.S., ed.: Advances in Kernel Methods: Support Vector Machines. MIT Press, Cambridge, MA (1998)