Available online at www.sciencedirect.com
SCIENCE DIRECT®
Neural Networks 18 (2005) 808–817
Neural Networks
www.elsevier.com/locate/neunet

2005 Special issue

Modelling ordinal relations with SVMs: An application to objective aesthetic evaluation of breast cancer conservative treatment☆

Jaime S. Cardoso[a,*], Joaquim F. Pinto da Costa[b], Maria J. Cardoso[c]

[a]*Faculdade de Engenharia, Universidade do Porto, INESC, Porto, Portugal*
[b]*Faculdade de Ciências, Universidade do Porto, Portugal*
[c]*Faculdade de Medicina, Universidade do Porto, Portugal*

Abstract

The cosmetic result is an important endpoint for breast cancer conservative treatment (BCCT), but the verification of this outcome remains without a standard. Objective assessment methods are preferred to overcome the drawbacks of subjective evaluation.

In this paper a novel algorithm is proposed, based on support vector machines, for the classification of ordinal categorical data. This classifier is then applied as a new methodology for the objective assessment of the aesthetic result of BCCT.

Based on the new classifier, a semi-objective score for quantification of the aesthetic results of BCCT was developed, allowing the discrimination of patients into four classes.

1. Introduction

There are two main goals in this work. The first is to develop a novel classification algorithm for ordinal data, in which there is a natural order among classes. The second is to apply that algorithm to the challenging problem of the objective evaluation of the aesthetic result of BCCT.

The first experimental results (Cardoso, da Costa, & Cardoso, 2005) to predict objectively the cosmetic outcome in BCCT, using a classifier based on support vector machines, showed promising performance. That led us to carry out a more exhaustive set of experiments, comparing several algorithms based on SVMs. The novel proposed algorithm attains the best generalization capability at the lowest computational load for the group of methods evaluated.

2. Classification methods for ordinal data

Many pattern recognition problems involve classifying examples into classes which have a natural ordering. Settings in which it is natural to rank instances arise in many fields, such as information retrieval (Herbrich, Graepel, & Obermayer, 1999a), collaborative filtering (Shashua & Levin, 2002) and econometric modeling (Mathieson, 1995).

Suppose that examples in a classification problem belong to one of k classes, numbered from 1 to k, corresponding to their natural order if one exists, and arbitrarily otherwise. The learning task is to select a prediction function $f(\mathbf{x})$ from a family of possible functions that minimizes the expected *loss*.

In the absence of reliable information on relative costs, a natural approach for unordered classes is to treat every misclassification as equally likely. This translates to adopting the non-metric indicator function $l_{0-1}(f(\mathbf{x}),y)=0$ if $f(\mathbf{x})=y$ and $l_{0-1}(f(\mathbf{x}),y)=1$ if $f(\mathbf{x})\neq y$, where $f(\mathbf{x})$ and y are the predicted and true classes, respectively. Measuring the performance of a classifier using the l_{0-1} loss function is equivalent to simply considering the misclassification error rate. However, for ordered classes, losses that increase with the absolute difference between the class numbers are more natural choices in the absence of better information (Mathieson, 1995).

☆ An abbreviated version of some portions of this article appeared in Cardoso, da Costa et al. (2005), published under the IEEE copyright.

* Corresponding author.
E-mail addresses: jaime.cardoso@inescporto.pt (J.S. Cardoso), jpcosta@fc.up.pt (J.F. Pinto da Costa), mjcard@med.up.pt (M.J. Cardoso).

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.06.023

The use of techniques designed specifically for ordered classes results in simpler classifiers, making it easier to interpret the factors that are being used to discriminate among classes (Mathieson, 1995). We propose a classifier for ordered classes based on support vector machines by reducing the problem of classifying ordered classes to the standard two-class problem.

2.1. The ABC of support vector machines

Consider briefly how the SVM binary pattern recognition problem is formulated (Vapnik, 1998).[1]

Assume the training set $\{\mathbf{x}_i^{(j)}\}$, where $j=1,2$, denotes the class number, $i=1,\ldots,\ell_j$ is the index within each class and $\mathbf{x}_i^{(j)} \in \mathbb{R}^p$, with p the dimension of the feature space.

For two training classes linearly separable in the selected feature space, the distinctive idea of SVM is to define a linear discriminant function $g(\mathbf{x})=\mathbf{w}^t\mathbf{x}+b$ in the feature space bisecting the two training classes and characterized by $g(\mathbf{x})=0$. However, there may be infinitely many such surfaces. To select the surface best suited to the task, the SVM maximizes the distance between the decision surface and those training points lying closest to it (the support vectors). It is easy to show (Vapnik, 1998) that maximizing this distance is equivalent to solving

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^t\mathbf{w} \\ \text{s.t.} \quad & -(\mathbf{w}^t\mathbf{x}_i^{(1)}+b) \geq +1 \quad i=1,\ldots,\ell_1 \\ & +(\mathbf{w}^t\mathbf{x}_i^{(2)}+b) \geq +1 \quad i=1,\ldots,\ell_2 \end{aligned} \tag{1}$$

If the training classes are not linearly separable in feature space, the inequalities in (1) can be relaxed using slack variables and the cost function modified to penalise any failure to meet the original (strict) inequalities. The problem becomes

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\mathbf{w}^t\mathbf{w} + C\sum_{j=1}^{2}\sum_{i=1}^{\ell_j}\mathrm{sgn}(\xi_i^{(j)}) \\ & -(w^t x_i^{(1)}+b) \geq +1-\xi_i^{(1)} \quad i=1,\ldots,\ell_1 \\ \text{s.t.} \quad & +(w^t x_i^{(2)}+b) \geq +1-\xi_i^{(2)} \quad i=1,\ldots,\ell_2 \\ & \xi_i^{(j)} \geq 0 \end{aligned} \tag{2}$$

The constraint parameter C controls the tradeoff between the dual objectives of maximizing the margin of separation and minimizing the misclassification error. For an error to occur, the corresponding ξ_i must exceed unity so $\sum_{j=1}^{2}\sum_{i=1}^{\ell_j}\mathrm{sgn}(\xi_i^{(j)})$ is an upper bound on the number of the training errors, that is $\sum l_{0-1}(f(\mathbf{x}_i^{(j)}),j)$, where $f(\mathbf{x}_i^{(j)})$ is the classification rule induced by the hyperplane $\mathbf{w}^t\mathbf{x}+b$. Hence the added penalty component is a natural way to assign an extra cost for errors.

However, optimization of the above is difficult since it involves a discontinuous function sgn (). As it is common in such cases, we choose to optimize a closely related cost function, and the goal becomes to

$$\min_{w,b,\xi_i} \frac{1}{2}\mathbf{w}^t\mathbf{w} + C\sum_{j=1}^{2}\sum_{i=1}^{\ell_j}\xi_i^{(j)} \tag{3}$$

under the same set of constraints as (2).

In order to account for different misclassification costs or sampling bias, the model can be extended to penalize the slack variables according to different weights in the objective function (Lin, Lee & Wahba, 2002):

$$\min_{w,b,\xi_i} \frac{1}{2}\mathbf{w}^t\mathbf{w} + \sum_{j=1}^{2}\sum_{i=1}^{\ell_j}C_i\xi_i^{(j)} \tag{4}$$

2.2. Modelling ordinal relations with SVMs

Let us formulate the problem of separating k ordered classes $C_1,\ldots,C_k$ in the spirit of SVMs. Consider the training set $\{\mathbf{x}_i^{(j)}\}$, where $j=1,\ldots,k$, denotes the class number, $i=1,\ldots_k\ell_j$ is the index within each class, and $\mathbf{x}_i^{(j)} \in \mathbb{R}^p$. Let $\ell=\sum_{j=1}\ell_j$ be the total number of training examples.

A risk functional that takes into account the ordering of the classes can be defined as

$$R(f) = \mathbf{E}[l^s(f(\mathbf{x}^{(j)},j)] \tag{5}$$

with

$$l^s(f(\mathbf{x}^{(j)}),j) = \min(|f(x^{(j)})-j|,s)$$

The empirical risk is the average of the number of mistakes, where the magnitude of a mistake is related to the total ordering:

$$R^s_{\mathrm{emp}}(f) = \frac{1}{\ell}\sum_{j=1}^{k}\sum_{i=1}^{\ell_j} l^s\left(f(\mathbf{x}_i^{(j)}),j\right).$$

Arguing as Herbrich, Graepel et al. (1999), we see that the role of parameter s (bounding the loss incurred in each example) is to allow for an incorporation of a priori knowledge about the probability of the classes, conditioned by $\mathbf{x}$, $P(j|\mathbf{x})$. This can be treated as an assumption on the concentration of the probability around a 'true' rank.

Designing a linear discriminator between two classes translates to choosing a weighted sum of the individual features and a bias to differentiate classes. Suppose that in the design of a k-class classifier all boundaries were constrained to have the same orientation—to choose the same weighted sum for all decisions. With this constraint

[1] The following introduction to SVMs is based largely on (Shilton, Ralph & Tsoi 2005).

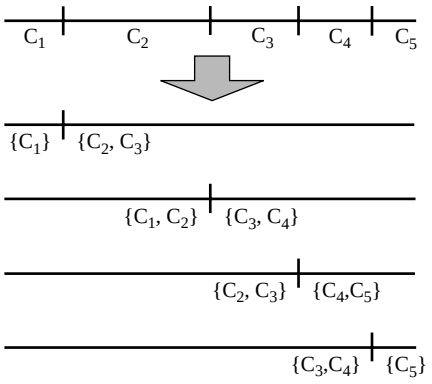


Fig. 1. Classes involved in the hyperplanes constraints, for $k=5$, $s=2$.

emerges a monotonic model, where a higher value in an attribute does not lead to a lower decision class. The output of the classifier would be just a set of weights, one for each feature, and a set of biases, the *scale* in the weighted sum. This simplified model captures better the essence of the problem. Another strength of this approach is the reduced number of parameters to estimate, which may lead to a more robust classifier, with greater capacity for generalization.

This rationale leads to a straight-forward generalization of the two-class separating hyperplane (Shashua & Levin, 2002). Define $k-1$ separating hyperplanes that separate the training data into k ordered classes by modeling the ranks as intervals on the real line—an idea with roots in the classical cumulative model, (Herbrich, Graepel et al., 1999; McCullagh & Nelder, 1989). The geometric interpretation of this approach is to look for $k-1$ parallel hyperplanes represented by vector $\mathbf{w}\in\mathbb{R}^p$ and scalars $b_1,\ldots,b_{k-1}$, such that the feature space is divided into equally ranked regions by the decision boundaries $\mathbf{w}^t\mathbf{x}+b_r$ $r\in\overline{1,k-1}$.

Going for a strategy to maximize the margin of the closest pair of classes, the goal becomes to maximize $\min|\mathbf{w}^t\mathbf{x}+b_i|/\|\mathbf{w}\|$. Recalling that an algebraic measure of the distance of a point to the hyperplane $\mathbf{w}^t\mathbf{x}+b$ is given by $(\mathbf{w}^t\mathbf{x}+b)/\|\mathbf{w}\|$, we can scale $\mathbf{w}$ and b_i so that the value of the minimum margin is $2/\|\mathbf{w}\|$.

The constraints to consider result from the $k-1$ binary classifications related to each hyperplane; the number of classes involved in each binary classification can be made dependent on a parameter s, Fig. 1. For the hyperplane $q\in\overline{1,k-1}$, the constraints result as

$$\begin{aligned}
&-(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_q)\geq+1 \quad j=\max(1,q-s+1),\ldots,q\\
&+(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_q)\geq+1 \quad j=q+1,\ldots,\min(k,q+s)
\end{aligned}$$

Our model can now be summarized as:

$$\min_{\mathbf{w},\mathbf{b}_i}\frac{1}{2}\mathbf{w}^t\mathbf{w}$$

s.t.

$$\begin{aligned}
&-(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_1)\geq+1 && j=1\\
&+(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_1)\geq+1 && j=2,\ldots,\min(k,1+s)\\
&\quad\vdots\\
&-(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_q)\geq+1 && j=\max(1,q-s+1),\ldots,q\\
&+(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_q)\geq+1 && j=q+1,\ldots,\min(k,q+s)\\
&\quad\vdots\\
&-(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_{k-1})\geq+1 && j=\max(1,k-s),\ldots,k-1\\
&-(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_{k-1})\geq+1 && j=k\\
&\quad\xi_i^{(j)}\geq 0
\end{aligned}\tag{6}$$

Reasoning as in the two-class SVM for the non-linearly separable dataset, the model becomes

$$\min_{\mathbf{w},\mathbf{b}_i,\xi_i}\frac{1}{2}\mathbf{w}^t\mathbf{w}+C\sum_{q=1}^{k-1}\sum_{j=\max(1,q-s+1)}^{\min(k,q+s)}\sum_{i=1}^{l_j}\operatorname{sgn}(\xi_{i,q}^{(j)})\tag{7}$$

s.t.

$$\begin{aligned}
&-(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_1)\geq+1-\xi_{i,1}^{(j)} && j=1\\
&+(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_1)\geq+1-\xi_{i,1}^{(j)} && j=2,\ldots,\min(k,1+s)\\
&\quad\vdots\\
&-(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_q)\geq+1-\xi_{i,q}^{(j)} && j=\max(1,q-s+1),\ldots,q\\
&+(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_q)\geq+1-\xi_{i,q}^{(j)} && j=q+1,\ldots,\min(k,q+s)\\
&\quad\vdots\\
&-(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_{k-1})\geq+1-\xi_{i,k-1}^{(j)} && j=\max(1,k-s),\ldots,k-1\\
&+(\mathbf{w}^t\mathbf{x}_i^{(j)}+b_{k-1})\geq+1-\xi_{i,k-1}^{(j)} && j=k\\
&\quad\xi_{i,q}^{(j)}\geq 0
\end{aligned}$$

Since each point $\mathbf{x}_i^{(j)}$ is involved $2.s$ times in the definition of the constraints, it can be shown to be misclassified $\min(|f(\mathbf{x}_i^{(j)})-j|,s)=l^s(f(\mathbf{x}_i^{(j)}),j)$ times, where $f(\mathbf{x}_i^{(j)})$ is the class estimated by the model. As for the two-class example, $\sum_{q=1}^{k-1}\sum_{j=\max(1,q-s+1)}^{\min(k,q+s)}\sum_{i=1}^{\ell_j}\operatorname{sgn}(\xi_{i,q}^{(j)})$ is an upperbound of $\sum_j\sum_i l^s(f(\mathbf{x}_i^{(j)}),j)$, proportional to the empirical risk.

Continuing the parallelism with the two-class SVM, the function to minimize simplifies to

$$\min_{\mathbf{w},\mathbf{b}_i,\xi_i}\frac{1}{2}\mathbf{w}^t\mathbf{w}+C\sum_{q=1}^{k-1}\sum_{j=\max(1,q-s+1)}^{\min(k,q+s)}\sum_{i=1}^{\ell_j}\xi_{i,q}^{(j)}\tag{8}$$

subject to the same constraints as (7).

As easily seen, the proposed formulation resembles the fixed margin strategy in (Shashua & Levin, 2002). However, instead of using only the two closest classes in the

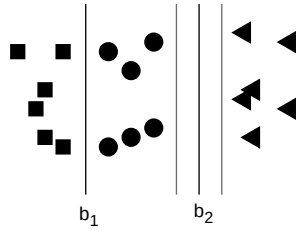
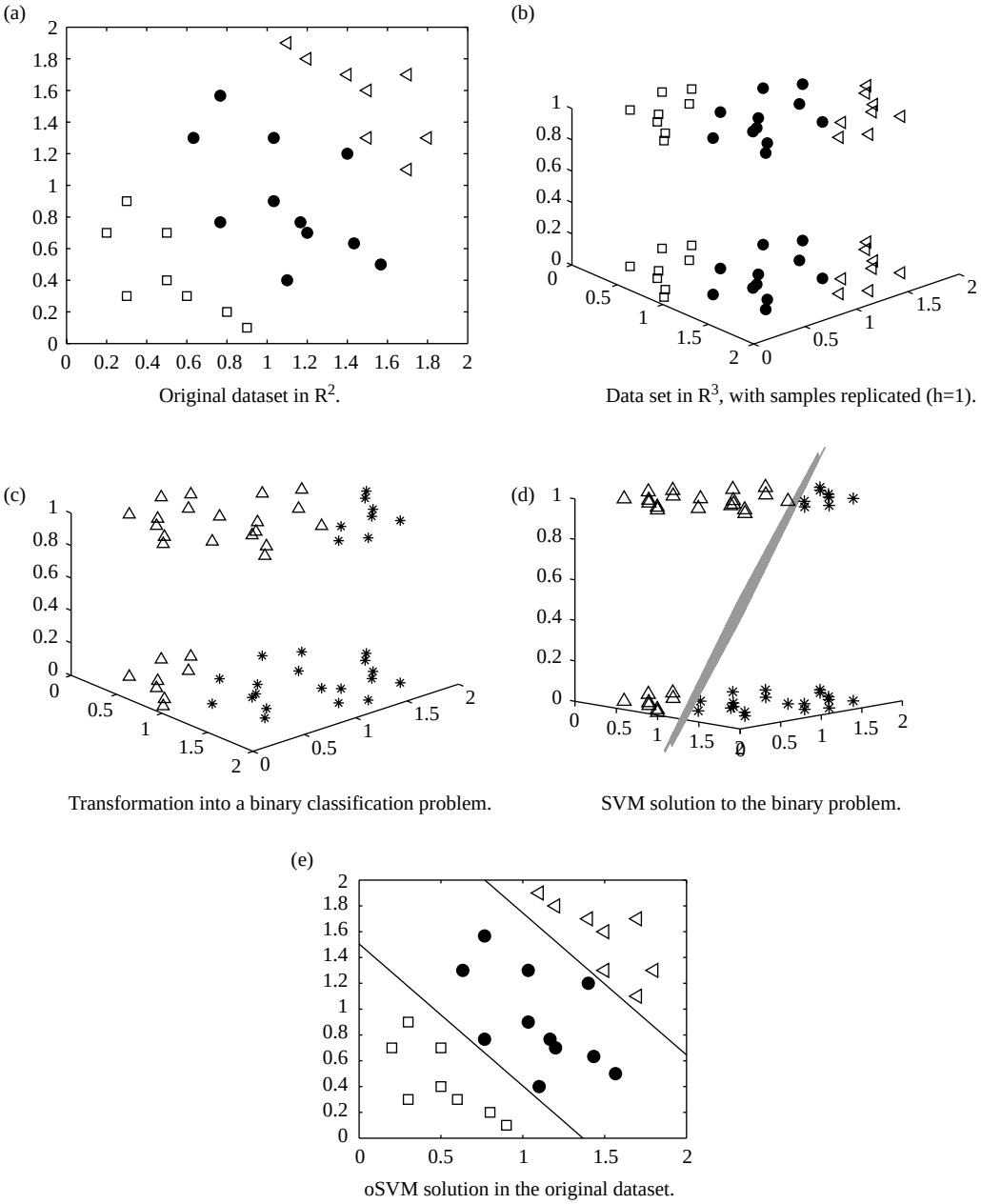


Fig. 2. Scalar b_2 is undetermined over an interval under the fixed margin strategy.

constraints of an hyperplane-more appropriate for the loss function $l_{0-1}()$, we adopt a formulation that captures better the performance of a classifier for ordinal data.

Two issues were identified in the above formulation. First, this is an incompletely specified model because the scalars b_i are not well defined. In fact, although the direction of the hyperplanes $\mathbf{w}$ is unique under the above formulation (proceeding as Vapnik (1998) for the binary case), the scalars $b_1,\ldots,b_{k-1}$ are not uniquely defined, Fig. 2.

Another issue is that, although the formulation was constructed from the two-class SVM, it is no longer solvable with the same algorithms. It would be interesting to accommodate this formulation under the two-class problem. That would allow the use of mature and optimized algorithms, developed for the training of support vector machines (Platt, 1998; Dong, Krzyzak & Suen, 2005).



Original dataset in R^2.

Data set in R^3, with samples replicated (h=1).

Transformation into a binary classification problem.

SVM solution to the binary problem.

oSVM solution in the original dataset.

Fig. 3. Proposed oSVM model in a toy example.

2.3. oSVM algorithm

To outline the rationale behind the proposed model, consider first an hypothetical, simplified scenario with three classes in $\mathbb{R}^2$. The plot of the dataset is presented in Fig. 3(a).

Using a transformation from the $\mathbb{R}^2$ initial feature-space to a $\mathbb{R}^3$ feature space, replicate each sample with different values in the new dimension, according to (Fig. 3(b)):

$$\mathbf{x} \in \mathbb{R}^2 \rightarrow \begin{bmatrix} \mathbf{x} \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{x} \\ h \end{bmatrix} \in \mathbb{R}^3, \quad \text{where h} = \text{const} \in \mathbb{R}$$

Define a binary training set in the high-dimensional space according to (Fig. 3(c)):

$$\begin{bmatrix} \mathbf{x}_i^{(1)} \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_i^{(1)} \\ h \end{bmatrix}, \begin{bmatrix} \mathbf{x}_i^{(2)} \\ h \end{bmatrix} \in \bar{C}_1$$

$$\begin{bmatrix} \mathbf{x}_i^{(2)} \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_i^{(3)} \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_i^{(3)} \\ h \end{bmatrix} \in \bar{C}_2$$

The binary SVM formulation for this extended training set can now be described as (with $\bar{\mathbf{w}} = [\mathbf{w}\, w_3]$, $\mathbf{w} \in \mathbb{R}^2$)

$$\begin{aligned} \min_{\bar{\mathbf{w}},b} \quad & \frac{1}{2}\bar{\mathbf{w}}^t\bar{\mathbf{w}} \\ s.t. \quad & -\left(\bar{\mathbf{w}}^t \begin{bmatrix} \mathbf{x}_i^{(1)} \\ 0 \end{bmatrix} + b\right) \geq +1 \\ & +\left(\bar{\mathbf{w}}^t \begin{bmatrix} \mathbf{x}_i^{(2)} \\ 0 \end{bmatrix} + b\right) \geq +1 \\ & +\left(\bar{\mathbf{w}}^t \begin{bmatrix} \mathbf{x}_i^{(3)} \\ 0 \end{bmatrix} + b\right) \geq +1 \\ & -\left(\bar{\mathbf{w}}^t \begin{bmatrix} \mathbf{x}_i^{(1)} \\ h \end{bmatrix} + b\right) \geq +1 \\ & -\left(\bar{\mathbf{w}}^t \begin{bmatrix} \mathbf{x}_i^{(2)} \\ h \end{bmatrix} + b\right) \geq +1 \\ & +\left(\bar{\mathbf{w}}^t \begin{bmatrix} \mathbf{x}_i^{(3)} \\ h \end{bmatrix} + b\right) \geq +1 \end{aligned} \tag{9}$$

But because

$$\begin{cases} \bar{\mathbf{w}}^t \begin{bmatrix} \mathbf{x}_i \\ 0 \end{bmatrix} = \mathbf{w}^t\mathbf{x}_i \\ \bar{\mathbf{w}}^t \begin{bmatrix} \mathbf{x}_i \\ h \end{bmatrix} = \mathbf{w}^t\mathbf{x}_i + w_3 h \end{cases},$$

and renaming b to b_1 and $b + w_3 h$ to b_2 the formulation above simplifies to

$$\begin{aligned} \min_{\mathbf{w},b_1,b_2} \quad & \frac{1}{2}\mathbf{w}^t\mathbf{w} + \frac{1}{2}\frac{(b_2 - b_1)^2}{h^2} \\ & -(\mathbf{w}^t\mathbf{x}_i^{(1)} + b_1) \geq +1 \\ & +(\mathbf{w}^t\mathbf{x}_i^{(2)} + b_1) \geq +1 \\ & +(\mathbf{w}^t\mathbf{x}_i^{(3)} + b_1) \geq +1 \\ \text{s.t.} \quad & -(\mathbf{w}^t\mathbf{x}_i^{(1)} + b_2) \geq +1 \\ & -(\mathbf{w}^t\mathbf{x}_i^{(2)} + b_2) \geq +1 \\ & +(\mathbf{w}^t x_i^{(3)} + b_2) \geq +1 \end{aligned} \tag{10}$$

Two points are worth to mention: (a) this formulation, being the result of a pure SVM method, has an unique solution (Vapnik, 1998); (b) this formulation equals the formulation (8) for ordinal data previously introduced, with $k = 3$, $s = k - 1 = 2$, and a slightly modified objective function by the introduction of a regularization member, proportional to the distance between the hyperplanes. The oSVM solution is the one that simultaneously minimizes the distance between boundaries and maximizes the minimum of the margins—Fig. 4. The h parameter controls the tradeoff between the objectives of maximizing the margin of separation and minimizing the distance between the hyperplanes.

To reiterate, this data extension method enabled us to formulate the classification of ordinal data as a standard SVM problem, removing the ambiguity in the solution by the introduction of a regularization term in the objective function.

With the material on how to construct a set of optimal hyperplanes for the toy example, we are now in a position to

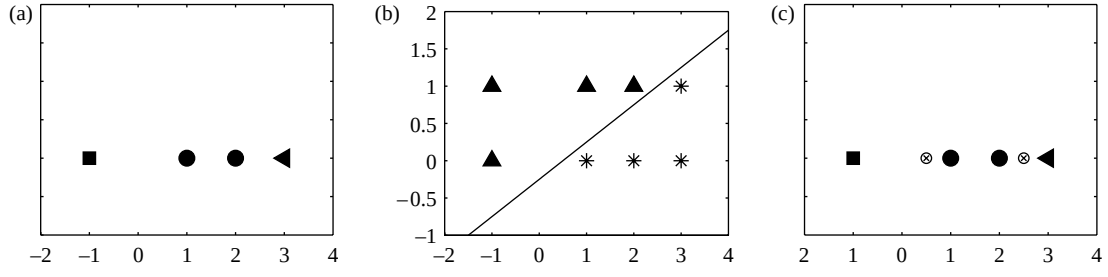


Fig. 4. Effect of the regularization member in the oSVM solution.

formally describe the construction of a support vector machine for ordinal regression. Define $\mathbf{0}$ as the sequence of $k-2$ zeros and $\mathbf{e}_q$ as the sequence of $k-2$ symbols 0,…,0, h, 0,…,0, with h in the q-th position. Considering the problem of separating k classes $C_1,\ldots C_k$ with training set $\{\mathbf{x}_i^{(j)}\}$, define a new high-dimensional binary training dataset as

$$\begin{bmatrix} \mathbf{x}_i^{(j)} \\ 0 \end{bmatrix} \in \begin{cases} \bar{C}_1 & j=1 \\ \bar{C}_2 & j=2,\ldots,\min(k,1+s) \end{cases}$$
$$\vdots$$
$$\begin{bmatrix} \mathbf{x}_i^{(j)} \\ \mathbf{e}_{q-1} \end{bmatrix} \in \begin{cases} \bar{C}_1 & j=\max(1,q-s+1),\ldots,q \\ \bar{C}_2 & j=q+1,\ldots,\min(k,q+s) \end{cases}$$
$$\vdots$$
$$\begin{bmatrix} \mathbf{x}_i^{(j)} \\ \mathbf{e}_{k-2} \end{bmatrix} \in \begin{cases} \bar{C}_1 & j=\max(1,k-1-s+1),\ldots,k-1 \\ \bar{C}_2 & j=k \end{cases}$$

After the simplifications and change of variables suggested in the toy example, the binary SVM formulation for this extended dataset yields

$$\min_{\mathbf{w},b_i,\xi_i} \frac{1}{2}\mathbf{w}^t\mathbf{w} + \frac{1}{h^2}\sum_{i=2}^{k-1}\frac{(b_i-b_1)^2}{2} + C \times \sum_{q=1}^{k-1}\sum_{j=\max(1,q-s+1)}^{\min(k,q+s)}\sum_{i=1}^{\ell_i}\xi_{i,q}^{(j)} \tag{11}$$

with the same set of constraints as (7).

This formulation for the high-dimensional dataset matches the proposed formulation for ordinal data up to an additional regularization member in the objective function. This additional member is responsible for the unique determination of the biases.

The scalars b_i yielded by the oSVM model can be further refined. Although the minimum margin principle only uniquely defines the bias for the closest pair of classes, it would be desirable to put the other hyperplanes as far as possible from the training examples. As such, a post-optimization operation, performing a one-dimensional SVM may be carried out, after the computation of the direction $\mathbf{w}$.

It is important to stress that the complexity of the SVM model does not depend on the dimensionality of the data. So, the only increase in the complexity of the problem is due to the duplication of the data (more generally, for a k-class problem, the dataset is increased ($2s$) times). As such, it compares favourably with the formulation in (Herbrich, Graepel, & Obermayer, 1999), which squares the dataset.

3. Aesthetic evaluation of breast cancer treatment

Breast cancer conservative treatment has been increasingly used over the last few years, as a consequence of its much more acceptable cosmetic outcome than traditional techniques, but with identical oncological results. Although considerable research has been put into BCCT techniques, diverse aesthetic results are common, highlighting the importance of this evaluation in institutions performing breast cancer treatment, so as to improve working practices.

Traditionally, aesthetic evaluation has been performed subjectively by one or more observers (Harris, Levene, Svensson & Hellman, 1979; Beadle, Silver, Botnick, Hellman, & Harris, 1984; Pierquin, Huart, Raynal, Otmezguine, Calitchi and Mazeron, 1991). However, this form of assessment has been shown to be poorly reproducible (Pezner, Patterson, Hill, Vora, Desai and Archambeau, 1985; Sacchini, Luini, Tana, Lozza, Galimberti and Merson, 1991; Sneeuw, Aaronson, Yarnold, Broderick, Ross and Goddard, 1992; Christie, O'Brien, Christie, Kron, Ferguson and Hamilton, 1996), which creates uncertainty when comparing results between studies. It has also been demonstrated that observers with different backgrounds evaluate cases in different ways (Cardoso, Santos, Cardoso, Barros & Oliveira, 2005).

Objective methods of evaluation have emerged as a way to overcome the poor reproducibility of subjective assessment and have until now consisted of measurements between identifiable points on patient photographs (Pezner, Patterson, Hill, Vora, Desai, Archambeau & Lipsett, 1985; Limbergen, Schueren & Tongelen, 1989; Christie et al., 1996). The correlation of objective measurements with subjective overall evaluation has been reported by several authors (Sacchini et al., 1991; Sneeuw et al., 1992; Christie et al., 1996; Al-Ghazal, Blamey, Stewart & Morgan, 1999). Until now though, the overall cosmetic outcome was simply the sum of the individual scores of subjective and objective individual indices (Noguchi, Saito, Mizukami, Nonomura, Ohta and Koyasaki, 1991; Sacchini et al., 1991; Sneeuw et al., 1992; Al-Ghazal et al., 1999).

The aim of this work was to develop an optimized semi-objective score for quantification of aesthetic results in BCCT, which would discriminate among four categories of overall subjective evaluation. This would constitute a useful indication for the development of a totally automatic program for the aesthetic evaluation of BCCT.

3.1. Data and method

Instead of heuristically weighting the individual indices in an overall measure, we introduced pattern classification techniques to find the correct contribution of each individual feature in the final result and the scale intervals for each class, constructing in this way an optimal rule to classify patients.

Reference classification: twenty-four clinicians working in twelve different countries were selected, based on their experience in BCCT (number of cases seen per year and/or participation in published work on evaluation of aesthetic results). They were asked to evaluate individually a series of 240 photographs taken from 60 women submitted to BCCT (surgery and radiotherapy). Treatment was completed at

least one year before the onset of the study and all patients signed an informed consent to participate. Photographs were taken (with a 4 M digital camera) in four positions with the patient standing on floor marks: facing, arms down; facing, arms up; left, side arms up; right, side arms up.

Participants were asked to evaluate overall aesthetic results, classifying each case into one of four categories: *excellent*—treated breast nearly identical to untreated breast; *good*—treated breast slightly different from untreated; *fair*—treated breast clearly different from untreated but not seriously distorted; *poor*—treated breast seriously distorted (Harris et al., 1979).

In order to obtain a consensus among observers, the Delphi process was used (Jones & Hunter, 1995; Hasson, Keeney & McKenna, 2000). Evaluation of each case was considered consensual when more than 50% of observers provided the same classification. When this did not occur, another round of agreement between observers was performed. For this, feedback sheets of all observers' results were sent by e-mail, disclosing consensual cases and asking for a revised opinion on non-consensual ones. Consensus in the first round occurred in 46/60 cases (77%). In the second round, answers were obtained from 22 of the 24 participants and consensus was reached in 59/60 cases (98%). See (Cardoso, Cardoso, Santos, Barros & Oliveira, 2005) for more details.

By means of the Delphi process each and every patient was classified in one of the four categories (Table 1): *poor*, *fair*, *good*, and *excellent*.

The evaluation of two individual aesthetic characteristics, scar visibility and colour dissimilarities between the breasts, were asked from the panel, using the same grading scale: *excellent*; *good*; *fair*; *poor*.

Feature selection: as possible objective features we considered those already identified by domain experts as relevant to the aesthetic evaluation of the surgical procedure (Pezner et al., 1985; Limbergen et al., 1989). The cosmetic result after breast conserving treatment is mainly determined by visible skin alterations or changes in breast volume or shape. Skin changes can consist of a disturbing surgical scar or radiation-induced pigmentation or telangiectasia (Limbergen et al., 1989). Breast asymmetry was assessed by Breast Retraction Assessment (BRA), Lower Breast Contour (LBC) or Upward Nipple Retraction (UNR)—Fig. 5. Because breast asymmetry was insufficient to discriminate among patients, we adopted the mean of the scar visibility and skin colour change, as measured by the Delphi panel, as additional features to help in the separation task, as we had not yet established the evaluation of those features by quantitative methods (Cardoso, da Costa et al., 2005).

Table 1
Distribution of patients over the four classes

Class	# cases
Poor	7
Fair	12
Good	32
Excellent	9

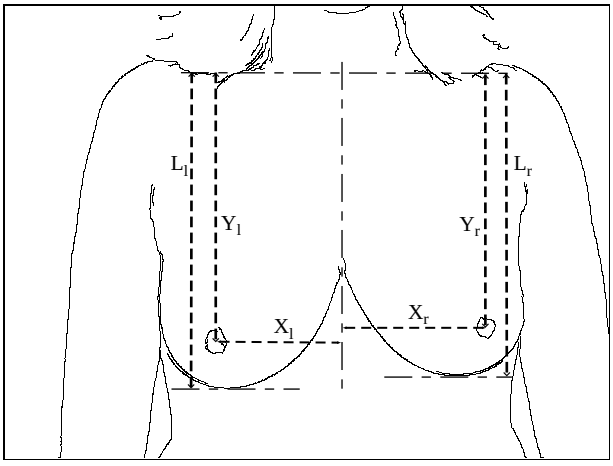


Fig. 5. LBC$=|L_r-L_l|$, BRA$=\sqrt{(X_r-X_l)^2+(Y_r-Y_l)^2}$, UNR$=|Y_r-Y_l|$.

Classifier: the *leave one out* method (Duda, Hart & Stork, 2001) was selected for the validation of the classifiers: the classifier is trained in a round-robin fashion, each time using the available dataset from which a single patient has been deleted; each resulting classifier is then tested on the single deleted patient.

When in possession of a *nearly separable* dataset, a simple linear separator is bound to misclassify some points. But the real question is if the *non-linearly-separable* data indicates some intrinsic property of the problem (in which case a more complex classifier, allowing more general boundaries between classes may be more appropriate) or if it can be interpreted as the result of *noisy points* (measurement errors, uncertainty in class membership, etc), in which case keeping the linear separator and accept some errors is more natural. Supported by Occam's razor principle ('one should not increase, beyond what is necessary, the number of entities required to explain anything'), the latter was the option taken in this research.

4. Experimental results

In this section we present experimental results for several SVM based algorithms applied to the prediction of the cosmetic result of BCCT. We compare the proposed oSVM with the following algorithms:

- Herbrich, Graepel et al. (1999) model, based on the correspondence of the ordinal regression task and the task of learning a preference relation on pairs of objects. A function loss was defined on pairs of objects and the classification task formulated in this space. The size of the new training set, derived from an ℓ-sized training set, can be as high as ℓ^2. Only the direction $\mathbf{w}$ was computed

directly from this model. Scalars b_i were obtained in a second step, performing a one-dimensional SVM.
- Shashua and Levin (2002) fixed margin model, in which the margin of the closest neighbouring classes is being maximized. This model resembles the proposed oSVM algorithm, as discussed previously.
- a generic multi-class SVM formulation, as described in (Franc & Hlaváč, 2002).
- pairwise SVM: because the classes are ordered, we performed three independent binary classifications, (*Poor*/*Fair*, *Fair*/*Good* and *Good*/*Excellent*), which can be interpreted as a simplification of the approach suggested by Hastie & Tibshirani (1998).

The h parameter of the oSVM algorithm was set to 100. No second step was performed to optimize the computation of the scalars b_i. Experiments were carried out in Matlab, using the Support Vector Machine toolbox, version 2.51, by Anton Schwaighofer. This toolbox was used to construct the oSVM classifier, the Herbrich, Graepel et al. (1999) model and the pairwise SVM. The Shashua & Levin (2002) fixed margin model was implemented making use of the MATLAB function `quadprog`. It was also used the STPRtool, version 2.01, for the implementation of the generic multi-class SVM. All the source code, as well as the datasets used, is available upon request to the authors.

The hypotheses the proposed algorithm share with the other specific algorithms a good generalization behaviour (and better than the multi-class and pairwise SVM), but presenting the least complexity.

4.1. Datasets

A fast visual checking of the quality of the data (Fig. 6(a)) shows that there is a data value that is logically inconsistent with the others: an individual (patient #31) labeled as *good* when in fact it is placed between *fair* and *poor* in the feature space. The classifiers were evaluated using datasets with and without this outlier in order to assess the behaviour in the presence of noisy examples.

In summary, results are reported for six different datasets: {*LBC* (arms down); *scar* visibility (mean); *skin* colour change (mean)}, {*BRA* (arms down); *scar* visibility (mean); *skin* colour change (mean)}, {*UNR* (arms down); *scar* visibility (mean); *skin* colour change (mean)}, each with 59 and 60 examples. In (Cardoso, da Costa et al., 2005) other datasets were evaluated, showing similar behaviour.

4.2. Results

Table 2 summarizes the generalization error estimated for each classifier, with the C parameter set to 10. The training time (normalized) for each classifier is also presented in Table 2. As seen, the proposed oSVM is much more efficient than the other specific algorithms, with the same performance. It is also apparent that algorithms specially designed for ordinal data perform better than generic algorithms for nominal classes. It is also noticeable the superiority of the LBC measure over the other asymmetry measures under study to discriminate classes.

Table 2
Average of generalization error and of training time

	LBC 59/60	BRA 59/60	UNR 59/60	Time
OSVM	0.09/0.14	0.16/0.19	0.18/0.19	1
Shashua	0.09/0.14	0.16/0.19	0.18/0.21	40
Herbrich	0.09/0.09	0.18/0.21	0.21/0.17	200
Multiclass	0.18/0.17	0.19/0.19	0.19/0.19	1.7
Pairwise	0.14/0.16	0.19/0.21	0.21/0.19	1.4

Finally, Fig. 6(b) shows the parallel planes (boundaries) separating each pair of classes, for the (*LBC*, *arms down*/*Scar*/*Colour*) feature set. The decision rule to classify unseen patients, using the boundary planes outputted by the oSVM classifier, can be cleanly stated as:

If (0.33 LBC + 1.00 Scar + 0.31 Colour)

$$\begin{cases} <1.38 & \Rightarrow \text{Excellent} \\ >1.38 \wedge <2.52 & \Rightarrow \text{Good} \\ >2.52 \wedge <3.92 & \Rightarrow \text{Fair} \\ >3.92 & \Rightarrow \text{Poor} \end{cases}$$

5. Discussion

Establishing the ideal method for the evaluation of the aesthetic result of BCCT is a major endpoint of breast cancer treatment. However this ideal goal has never been considered to be attained. The subjective evaluation of the esthetic result of BCCT was and still is by far the more commonly used method. Preferably the evaluation would be done by a panel of observers (Sneeuw et al., 1992; Al-Ghazal et al., 1999) and the most popular criteria were developed by Harris et al. (1979) for four classes. However, the low reproducibility among observers remains one of the weakest points of this system. Our own previous study (Cardoso, Cardoso, 2005) reported on the evaluation of 55 cases by 13 observers (four experienced, four medium experienced and five inexperienced). Agreement was higher in the group of experienced observers ($k=0.59$) than in the medium experienced ($k=0.35$) and inexperienced observers ($k=0.33$). To overcome the probable low values of agreement that would always be obtained with a subjective classification, we organized a Delphi consensus panel (24 experts classifying 60 cases into four classes), trying to achieve a gold standard classification that would serve us to design our system. As others before, we started by correlating subjective and objective measures. Pezner et al. (1985) were the mentors of this type of evaluation, producing

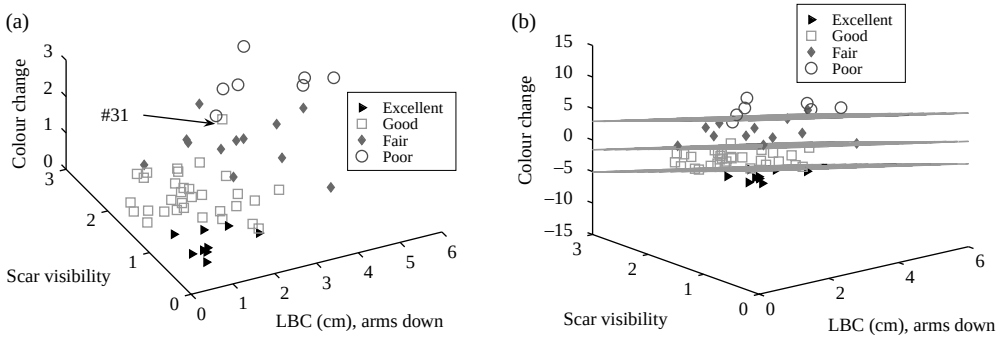


Fig. 6. (a) Three features space. (b) Decision boundaries, for the oSVM model.

a quantitative scoring system (Breast Retraction Assessment). Our work demonstrated that, in spite of a good correlation between the overall panel evaluation and the mean or median of the individual objective measurements, the wide dispersion of the measurements inside each class implies the need of a more principled approach. (Cardoso, da Costa et al., 2005). Limbergen et al. (1989) were defenders of a quantitative assessment but finally concluded that both, qualitative and quantitative, methods should be used together. Many others have followed this line of action (Borger & Keijser, 1987; Sacchini et al., 1991; Sneeuw et al., 1992; Christie et al., 1996). Trying to go a step further than Noguchi et al. (1991), who considered the overall result as sum of the individual indices, we tried to attain an optimized objective classification of patients. The considered objective measures for breast asymmetry (BRA, LBC and UNR) turned out to be insufficient to discriminate patients' classes (Cardoso, da Costa et al., 2005). Complementing the asymmetry information with the level of scar visibility and skin colour change, enabled us to successfully approach the problem as a pattern classification task for ordinal data.

Perhaps the main contribution, as far as pattern classification is concerned, is the development of a new technique to classify ordinal data, based on SVMs. This new method is likely to produce a simpler and more robust classifier, and compares favourably with state-of-the-art methods.

Using different kinds of classifiers, several feature sets were evaluated in terms of expected error rate over unseen cases, with the set LBC in arms down position, mean of scar visibility, and mean of skin colour change attaining the best value. It was also interesting to note that most of the misclassified cases are located in the transition between two classes in the feature space and correspond to individuals with a low agreement value in the answers of the Delphi panel.

We have presented a comprehensive and theoretically sound methodology to classify semi-objectively the esthetic result of BCCT. Other directions for future work include the definition of objective measures for the scar visibility and skin colour changes and automation of the evaluation process, developing specific software for that purpose.

Acknowledgements

The authors would like to thank Prof. Ayres-de-Campos for his help in preparation of the manuscript.

References

Al-Ghazal, S. K., Blamey, R. W., Stewart, J., & Morgan, A. L. (1999). The cosmetic outcome in early breast cancer treated with breast conservation. *European Journal of Surgical Oncology*, *25*, 566–570.

Beadle, G. F., Silver, B., Botnick, L., Hellman, S., & Harris, J. R. (1984). Cosmetic results following primary radiation therapy for early breast cancer. *Cancer*, *54*, 2911–2918.

Borger, J. H., & Keijser, A. H. (1987). Conservative breast cancer treatment: Analysis of cosmetic results and the role of concomitant adjuvant chemotherapy. *International Journal of Radiation Oncology Biology Physics*, *13*, 1173–1177.

Cardoso, M. J., Cardoso, J. S., Santos, A. C., Barros, H., & Oliveira, M. C. (accepted for publication). Interobserver agreement and consensus over evaluation of breast cancer conservative treatment. *The Breast*doi: 10.1016/j.breast.2005.04.013.

Cardoso, J. S., da Costa, J. F. P., & Cardoso, M. J. (2005). SVMs applied to objective aesthetic evaluation of conservative breast cancer treatment. In *Proceedings of the International Joint Conference on Neural Networks, July 31–August 4 2005, Montreal, Canada*.

Cardoso, M. J., Santos, A. C., Cardoso, J. S., Barros, H., & Oliveira, M. C. (2005b). Choosing observers for evaluation of aesthetic results in breast cancer conservative treatment. *International Journal of Radiation Oncology, Biology and Physics*, *61*, 879–881.

Christie, D. R. H., O'Brien, M.-Y., Christie, J. A., Kron, T., Ferguson, S. A., Hamilton, C. S., et al. (1996). A comparison of methods of cosmetic assessment in breast conservation treatment. *Breast*, *5*, 358–367.

Dong, J., Krzyzak, A., & Suen, C. Y. (2005). Fast SVM training algorithm with decomposition on very large data sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(4), 603–618.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification*. Wiley Interscience.

Franc, V., & Hlavác, V. (2002). Multi-class support vector machine. In R. Kasturi, D. Laurendeau, & C. Suen, *16th International conference on pattern recognition* (Vol. 2), 236–239.

Harris, J. R., Levene, M. B., Svensson, G., & Hellman, S. (1979). Analysis of cosmetic results following primary radiation therapy for stages i and ii carcinoma of the breast. *International Journal of Radiation Oncology Biology Physics*, *5*, 257–261.

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the delphi survey technique. *Journal of Advanced Nursing*, *32*, 1008–1015.

Hastie, T., & Tibshirani, R. (1998). Classification by pairwise coupling. In M. I. Jordan, M. J. Kearns, & S. A. Solla, *Advances in neural information processing systems* (Vol. 10). The MIT Press.

Herbrich, R., Graepel, T., & Obermayer, K. (1999). Regression models for ordinal data: A machine learning approach. *Tech. Rep. TR-99/03*, TU Berlin.

Herbrich, R., Graepel, T.,& Obermayer, K. (1999). Support vector learning for ordinal regression. In *Ninth international conference on artificial neural networks ICANN*. (pp. 97–102) Vol. 1.

Jones, J., & Hunter, D. (1995). Consensus methods for medical and health services research. *British Medical Journal*, *311*, 376–380.

Limbergen, E. V., Schueren, E. V., & Tongelen, K. V. (1989). Cosmetic evaluation of breast conserving treatment for mammary cancer. 1. Proposal of a quantitative scoring system. *Radiotherapy and oncology*, *16*, 159–167.

Lin, Y., Lee, Y., & Wahba, G. (2002). Support vector machines for classification in nonstandard situations. *Machine Learning*, *46*, 191–202.

Mathieson, M. J. (1995). Ordinal models for neural networks. In A. Refenes, Y. Moody, & J. Moody (Eds.), *Neural networks in financial engineering*. Singapore: World Scientific.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. New York: Chapman and Hall.

Noguchi, M., Saito, Y., Mizukami, Y., Nonomura, A., Ohta, N., Koyasaki, N., et al. (1991). Breast deformity, its correction, and assessment of breast conserving surgery. *Breast Cancer Research and Treatment*, *18*, 111–118.

Pezner, R. D., Patterson, M. P., Hill, L. R., Vora, N., Desai, K. R., Archambeau, J. O., et al. (1985). Breast retraction assessment: An objective evaluation of cosmetic results of patients treated conservatively for breast cancer. *International Journal of Radiation Oncology Biology Physics*, *11*, 575–578.

Pierquin, B., Huart, J., Raynal, M., Otmezguine, Y., Calitchi, E., Mazeron, J. J., et al. (1991). Conservative treatment for breast cancer:Long-term results (15 years). *Radiotherapy Oncology*, *20*, 16–23.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods-support vector Learning*. (pp. 185–208).

Sacchini, V., Luini, A., Tana, S., Lozza, L., Galimberti, V., Merson, M., et al. (1991). Quantitative and qualitative cosmetic evaluation after conservative treatment for breast cancer. *European Journal Cancer*, *27*, 1395–1400.

Shashua, A., & Levin, A. (2002). Ranking with large margin principle: Two approaches. In *Neural information and processing systems* (NIPS).

Shilton, A., Ralph, M. P. D., & Tsoi, A. C. (2005). Incremental training of support vector machines. *IEEE Transactions on Neural Networks*, *16*(1), 114–131.

Sneeuw, K. C., Aaronson, N. K., Yarnold, J. R., Broderick, M., Ross, J. R. G., & Goddard, A. (1992). Cosmetic and functional outcomes of breast conserving treatment for early stage breast cancer. 1. Comparison of patients' ratings, observers' ratings, and objective assessments. *Radiotherapy and Oncology*, *25*, 153–159.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.