Knowledge-Based Systems 148 (2018) 41–46

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Redefining support vector machines with the ordered weighted average

Sebastián Maldonado[a,*], José Merigó[b], Jaime Miranda[b]

[a] Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Monseñor Álvaro del Portillo 12455, Las Condes, Santiago, Chile
[b] Department of Management Control and Information Systems, School of Economics and Business, University of Chile, Santiago, Chile

ARTICLE INFO

Article history:
Received 12 November 2017
Revised 12 January 2018
Accepted 10 February 2018
Available online 15 February 2018

Keywords:
OWA operators
OWA quantifiers
Support vector machines
Hinge loss

ABSTRACT

In this work, the classical soft-margin Support Vector Machine (SVM) formulation is redefined with the inclusion of an Ordered Weighted Averaging (OWA) operator. In particular, the hinge loss function is rewritten as a weighted sum of the slack variables to guarantee adequate model fit. The proposed two-step approach trains a soft-margin SVM first to obtain the slack variables, which are then used to induce the order for the OWA operator in a second SVM training. Originally developed as a linear method, our proposal extends it to nonlinear classification thanks to the use of Kernel functions. Experimental results show that the proposed method achieved the best overall performance compared with standard SVM and other well-known data mining methods in terms of predictive performance.

1. Introduction

Support Vector Machines (SVMs) [31] has gained popularity among researchers and practitioners thanks to its theoretical advantages, such as superior predictive performance, adequate generalization to new samples thanks to the Structural Risk Minimization (SRM) principle [32], and the absence of local minima via convex quadratic optimization [21,31]. Support Vector Machines has been successfully applied in various domains, including computer vision [3], medical diagnosis [27], bioinformatics [7,9,22], and document classification [45], among others.

When dealing with data, it is necessary many times to aggregate the information in order to provide a representative view. In the literature, there are many aggregation operators [4,16,42]. The Ordered Weighted Average (OWA) [36,40] is a very popular one. The OWA operator is an aggregation operator that provides a parameterized family of aggregation operators between the minimum and the maximum. From a decision-making point of view, the OWA allows decision makers to analyze the data according to their own optimistic or pessimistic attitudes. The OWA operator has been extended and generalized within a wide range of frameworks [12]. For example, Yager and Filev [39] presented the induced OWA (IOWA) operator, and that work was further extended by Merig'o and Gil-Lafuente [24] using generalized and quasi-arithmetic means. Other studies use OWA operators with distance measures [23,34], moving averages [25], utilities [15], Bonferroni means [5,38], interval numbers [35,43], and fuzzy information [44].

In this work, a novel SVM strategy based on OWA operators is introduced for binary classification. The idea is to replace the traditional hinge loss function by a weighted sum using an OWA operator, penalizing the classification errors unevenly according to their distance from the hyperplane. A two-step algorithm is proposed: First, the classical soft-margin SVM is trained, given the order of the samples in terms of their distance from the classifier as the relevant output for this step. Next, SVM is re-trained using an OWA operator, leading to a final classifier aimed at being more effective than the traditional SVM but with the same complexity. The method is first presented as a linear classification strategy, and subsequently extended as a Kernel method.

This paper is structured as follows: previous work on OWA operators and Support Vector Machines are discussed in Section 2. The proposed framework for SVM classification based on OWA operators is described in Section 3. In Section 4, experimental results using benchmark datasets are given. Finally, the main conclusions of this study are presented in Section 5.

2. Theoretical background on OWA operators and SVM classification

In this section, the concept of OWA operators is first introduced, discussing the relevant variants that are considered for the development of our proposal. Next, we describe the mathematical

* Corresponding author.
E-mail addresses: smaldonado@uandes.cl (S. Maldonado), jmerigo@fen.uchile.cl (J. Merigó), jmirandap@fen.uchile.cl (J. Miranda).

https://doi.org/10.1016/j.knosys.2018.02.025
0950-7051/© 2018 Elsevier B.V. All rights reserved.

derivation of the soft-margin SVM developed by Cortes and Vapnik [10], and its extension to nonlinear classification.

2.1. The ordered weighted average

The OWA operator [36] aggregates the information providing a parameterized family of aggregation operators between the minimum and the maximum. It is widely used when aggregating the data according to the attitudinal character of the decision maker. Note that the OWA operator does not assign a weight directly to each of the elements of a set. Instead, it weights the data according to the ranking they achieve inside the set when comparing their numerical values. This situation is very useful when there is no information about the degrees of importance of the weights. The OWA operator is defined as follows:

Definition 1. An OWA operator of dimension n is mapping $OWA: \mathbb{R}^n \rightarrow \mathbb{R}$ that has an associated weighting vector W of dimension n with $\sum_{j=1}^{n} w_j = 1$ and $w_j \in [0, 1]$, such that:

$$OWA(a_1, a_2, \ldots, a_n) = \sum_{j=1}^{n} w_j b_j \tag{1}$$

where b_j is the j_{th} largest of the a_i.

A wide range of possible aggregation operators can be obtained when varying the weighting vector. The following ones are worth noting among others [24,37]:

- If $w_1 = 1$ and $w_j = 0$ for all $j \neq 1$, the OWA operator becomes the maximum.
- If $w_n = 1$ and $w_j = 0$ for all $j \neq n$, we get the minimum.
- The median-OWA occurs under two different situations that depend on the size of n. If n is odd, it appears when $w_{(n+1)/2} = 1$ and $w_j = 0$ for all others. And if n is even, the median is not a single number and an additional method is required. A common one is to assign $w_{n/2} = w_{(n/2)+1} = 0.5$ and $w_j = 0$ for all others.
- If $w_j = 1/n$, for all j, the OWA becomes the classical arithmetic mean.
- The step-OWA operator appears when $w_k = 1$ and $w_j = 0$ for all $j \neq k$. Note that it includes the maximum, the minimum, and the odd median as particular cases.
- The olympic-OWA makes an average of all the weights except the first and the last one. That is, $w_1 = w_n = 0$, and for all others $w_j = 1/(n-2)$.
- The generalized S-OWA operator appears if $w_1 = (1/n)(1-(\alpha+\beta))+\alpha$, $w_n = (1/n)(1-(\alpha+\beta))+\beta$, and $w_j = (1/n)(1-(\alpha+\beta))$ for $j = 2$ to $n-1$, where $\alpha, \beta \in [0, 1]$ and $\alpha+\beta \leq 1$. Observe that if $\alpha = 0$, it becomes the 'and-like' S-OWA and if $\beta = 0$, the 'or-like' S-OWA.

There are many other approaches in the literature for obtaining the OWA weights [24,37]. A practical approach is to use linguistic quantifiers [19,36]. The weights are generated by using a regular increasing monotone (RIM) quantifier Q as follows:

$$w_j = Q\left(\frac{j}{n}\right) - Q\left(\frac{j-1}{n}\right) \qquad \forall\, j. \tag{2}$$

Note that the weights generated through this method accomplish $\sum_{j=1}^{n} w_j = 1$ and $w_j \in [0, 1]$. This approach can generate a wide range of weighting vectors. The following ones are used in this study [19]:

- Basic linguistic quantifier:

$$Q(r) = r^{\alpha} \qquad \alpha \geq 0. \tag{3}$$

Here, the weights are obtained by using:

$$w_j = \left(\frac{j}{n}\right)^{\alpha} - \left(\frac{j-1}{n}\right)^{\alpha} \qquad \forall\, j. \tag{4}$$

- Quadratic linguistic quantifier [28]:

$$Q_q(r) = \left(\frac{1}{1-\alpha(r)^{0.5}}\right) \qquad \alpha \geq 0, \tag{5}$$

and the weights are calculated in the following way:

$$w_j = \left(\frac{1}{1-\alpha\left(\frac{j}{n}\right)^{0.5}}\right) - \left(\frac{1}{1-\alpha\left(\frac{j-1}{n}\right)^{0.5}}\right) \qquad \forall\, j. \tag{6}$$

- Exponential linguistic quantifier:

$$Q_e(r) = e^{-\alpha r}. \tag{7}$$

The weights for this quantifier are obtained as follows:

$$w_j = e^{-\alpha\left(\frac{j}{n}\right)} - e^{-\alpha\left(\frac{j-1}{n}\right)} \qquad \forall\, j. \tag{8}$$

- Trigonometric linguistic quantifier:

$$Q_t(r) = \arcsin(r\alpha), \tag{9}$$

with the following formula for the weights:

$$w_j = \arcsin\left(\alpha\left(\frac{j}{n}\right)\right) - \arcsin\left(\alpha\left(\frac{j-1}{n}\right)\right) \qquad \forall\, j. \tag{10}$$

The OWA operator is monotonic, commutative, bounded, and idempotent. In order to characterize the aggregation, there are several measures including the degree of orness-andness and the entropy of dispersion [36]. The degree of orness is formulated as follows:

$$\alpha(W) = \sum_{j=1}^{n} w_j \frac{n-j}{n-1} \tag{11}$$

Note that the andness is the complement of the orness. That is, andness = 1-orness.

The entropy of dispersion follows the methodology of Shannon [30] and applies it onto the OWA operator as follows:

$$H(W) = -\sum_{j=1}^{n} w_j \ln(w_j) \tag{12}$$

Note that $H(W) = 0$ for the step-OWA and its particular cases while the maximum entropy appears for the arithmetic mean because the weights show the highest dispersion between them. Thus, $H(W) = \ln n$.

The OWA operator can be generalized by using generalized and quasi-arithmetic means forming the generalized OWA and the quasi-arithmetic OWA (Quasi-OWA) operator [13,24]. The Quasi-OWA operator is very similar to the OWA operator with the difference that we introduce a strictly continuous monotonic function. That is:

$$Quasi - OWA(a_1, a_2, \ldots, a_n) = g^{-1}\left(\sum_{j=1}^{n} w_j g(b_j)\right) \tag{13}$$

where $g(b)$ is a strictly continuous monotonic function.

Observe the following particular cases of the $Quasi - OWA$ operator:

- The OWA operator when $g(b) = b$.
- The quadratic OWA when $g(b) = b^2$.
- The geometric OWA if $g(b) \rightarrow b^0$.
- The cubic OWA when $g(b) = b^3$.
- The harmonic OWA if $g(b) = b^{-1}$.
- The generalized OWA when $g(b) = b^{\lambda}$.

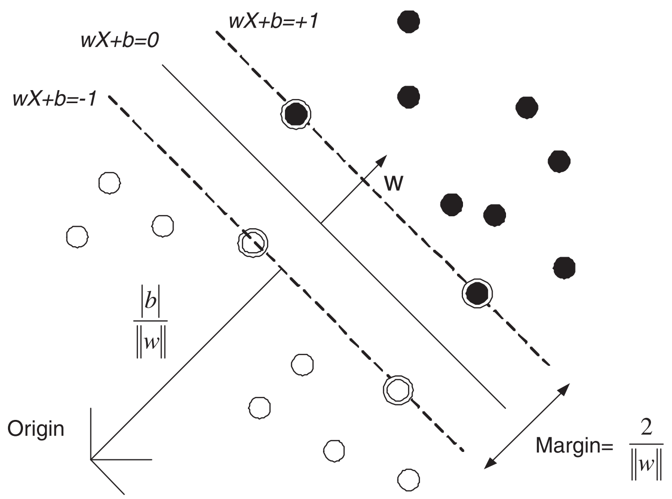

Fig. 1. Geometrical interpretation for the soft-margin SVM method. The support vectors are circled.

OWA operators have been applied in a wide range of fields [40]. Note that there are some studies that have considered the use of OWA operators in the context of SVM classification [1,18,41], but, unlike our proposal, they do not modify the SVM formulation. Alajlan et al. [1], for example, proposed an ensemble of SVM classifiers for pattern recognition on hyperspectral images, using OWA operators for combining the various model outputs.

2.2. Support vector machines

Consider the general binary classification problem of classifying m objects belonging to two sets denoted by $\mathcal{I}_1$ and $\mathcal{I}_2$. Each object i ($i = 1, \ldots, m$.) has n features which are stored in the i-th row of an $m \times n$ matrix $\mathbf{X}$. For each object i, y_i defines its label as follows: $y_i = 1$ if object i belongs to set $\mathcal{I}_1$ or $y_i = -1$ if object i belongs to set $\mathcal{I}_2$.

In broad terms, SVM constructs a separating hyperplane $\mathbf{w}^\top\mathbf{x} + b = 0$ that classifies the objects $\mathbf{x}_i$ correctly, while maximizing the separation margin between both classes $\frac{2}{\|\mathbf{w}\|}$ (see Fig. 1).

The margin maximization is theoretically motivated by the error bounds of the technique's generalization properties. Thus, the probability of misclassifying a new object is bounded by a function that decreases with the value of the margin [31]. The following quadratic programming (QP) problem is solved by SVM:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m, \\ & \xi_i \geq 0, \qquad i = 1, \ldots, m, \end{aligned} \tag{14}$$

where $C > 0$ is a parameter that controls the relative importance of the model fit over the margin maximization [31]. For a given training sample i, ξ_i represents how far it lies on the wrong side of the corresponding canonical hyperplane [31]. The canonical hyperplanes are the ones that 'support' the classifier, defining the boundaries for the margin [6] (the dashed lines in Fig. 1).

To build a non-linear decision surface, SVM maps the training points of dimension n onto a feature space of higher dimension using a projection function $\phi(\cdot)$. Thanks to the duality theory, SVM is able to construct a nonlinear classifier without the need of defining this projection function $\phi(\cdot)$. In the dual form of Formulation (14), the training examples appear only as scalar products, allowing the use of a Kernel function $K(\mathbf{x}_i, \mathbf{x}_s) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$. The Kernel-based version for the soft-margin SVM method follows [29]:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{m}\alpha_i - \tfrac{1}{2}\sum_{i,s=1}^{m}\alpha_i\alpha_s y_i y_s K(\mathbf{x}_i, \mathbf{x}_s) \\ \text{s.t.} \quad & \sum_{i=1}^{m}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m, \end{aligned} \tag{15}$$

where $\boldsymbol{\alpha}$ are the Lagrange multipliers related to the constraints in (14). Among the various Kernel functions, a frequent choice is the Gaussian, which has the following form:

$$K(\mathbf{x}_i, \mathbf{x}_s) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_s||^2}{2\sigma^2}\right), \tag{16}$$

where $\sigma > 0$ is the width parameter [29]. Formulation (15) can be solved efficiently, using the Sequential Minimal Optimization (SMO) technique [26], for example.

Finally, the decision rules for these two SVM formulations follow: given a new sample $\mathbf{x}^*$ with an unknown label, $y^* = sign(\mathbf{w}^\top\mathbf{x}^* + b)$ and $y^* = sign(\sum_{i' \in \mathcal{SV}}\alpha_{i'}y_{i'}K(\mathbf{x}_{i'}, \mathbf{x}^*) + b)$ for the linear and Kernel-based case, respectively [6], where $\mathcal{SV}$ is the set of support vectors, i.e. all samples from the training set with $\alpha_i > 0$.

3. The proposed framework for SVM classification based on OWA operators

The main idea is to modify the loss function used in SVM classification by incorporating an OWA operator. In order to do this, a two-step methodology is proposed: First, the soft-margin SVM method is applied (Formulation (14)) to obtain the distance between each training sample and its respective canonical hyperplane, given by $\boldsymbol{\xi}^\ddagger = 1 - y_i(\mathbf{w}^\top\mathbf{x}_i + b)$. Notice that this first SVM training serves only as an ordering method. Next, the soft-margin SVM formulation is redefined by including an OWA operator, instead of the hinge loss function, where the aggregation is ordered by the $\boldsymbol{\xi}$‡ values in ascending order.

The reasoning behind the proposed approach is to weight down correctly-classified samples, increasing the importance of boundary cases and misclassified objects in a second training process. This idea resembles adaptive methods like Adaboost, in which subsequent classifiers are tweaked in favor of those objects misclassified by previous learners in an iterative process [14]. Notice that $\boldsymbol{\xi}$‡ is similar to $\boldsymbol{\xi}$ when $\boldsymbol{\xi} > 0$; that is, a data point is either inside the margin or misclassified, but $\boldsymbol{\xi}$‡ can be negative, representing the distance of a correctly classified sample to its corresponding canonical hyperplane. The measure $\boldsymbol{\xi}$‡ is chosen over $\boldsymbol{\xi}$ for the OWA operator since it allows sorting all observations avoiding ties. It is important to keep in mind that most $\boldsymbol{\xi}$ values can be zero in datasets with little class overlap.

Formally, an OWA operator is defined by constructing a weighting vector $\mathbf{W}$, with $W_i \in (0, 1)$ $\forall i = 1, \ldots, m$ and $\sum_i W_i = 1$, using an OWA quantifier, or an alternative method from the literature (see Section 2). The quantifier parameter α can be set via cross-validation. Notice that the OWA weighting vector $\mathbf{W}$ is different from the solution vector $\mathbf{w}$ obtained by SVM.

The first SVM is trained using the standard hinge loss function, with $\boldsymbol{\xi}_1^\ddagger$ being the relevant output for this step. Then, the OWA operator $F(\boldsymbol{\xi}_1^\ddagger, \boldsymbol{\xi}_2, \mathbf{W}) = \mathbf{W}^\top\boldsymbol{\xi}_2^*$ is used in the objective function of the second SVM problem, where $\boldsymbol{\xi}_2^*$ is the vector consisting of $\boldsymbol{\xi}_2$ put in a descending order based on $\boldsymbol{\xi}_1^\ddagger$; that is, W_i is the weight associated with the ith largest value of $\boldsymbol{\xi}_1^\ddagger$. Following the idea behind OWA operators, weighting vector $\mathbf{W}$ is associated with a particular ordered position based on the $\boldsymbol{\xi}_1^\ddagger$ values rather than on a particular element. Therefore, the OWA operator F is used on a new set of slack variables $\boldsymbol{\xi}_2$, whose values are determined in a second SVM optimization process, but with the order induced by the values obtained in the first SVM model. Notice that each weight is re-ordered to match the corresponding sample from the first SVM training. The proposal is detailed in Algorithm (1) for the linear version.

Algorithm 1 OWA–SVM Algorithm, linear version.

```
Input: Training tuples $(\mathbf{x}_i, y_i)$, SVM soft-margin parameter $C$, OWA quantifier parameter $\alpha$.
Output: SVM classifier $(\mathbf{w}, b)$.

1. $\boldsymbol{\xi}_1^{\ddagger} \leftarrow$ soft-margin SVM training, Formulation (14), with parameter $C$.
2. $\mathbf{W} \leftarrow$ OWA quantifier based on quantifier parameter $\alpha$.
3. $(\mathbf{w}, b) \leftarrow$ OWA–SVM training:

   $\min_{\mathbf{w},b,\boldsymbol{\xi}_2} \quad \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{\overline{W}} \cdot F(\boldsymbol{\xi}_1^{\ddagger}, \boldsymbol{\xi}_2, \mathbf{W})$
   $\text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_{2,i}, \quad i = 1, \dots, m,$        (17)
   $\xi_{2,i} \geq 0, \quad i = 1, \dots, m.$

   where $\overline{W}$ is the mean value of vector $\mathbf{W}$. The idea behind this quotient is to normalize the effect of the weights $\mathbf{W}$.
```

Algorithm (1) can be extended easily to Kernel functions. The $\boldsymbol{\xi}\ddagger$ values can be computed from the solution of the dual SVM formulation: It follows from the derivation of the dual problem of Formulation (14) that $\mathbf{w}^\top \mathbf{x}_i$ becomes $\sum_{i' \in \mathcal{SV}} \alpha_{i'} y_{i'} K(\mathbf{x}_{i'}, \mathbf{x}_i)$ when Kernel functions are introduced [6]. Then,

$$\xi_i^{\ddagger} = 1 - y_i\left(\sum_{i' \in \mathcal{SV}} \alpha_{i'} y_{i'} K(\mathbf{x}_{i'}, \mathbf{x}_i) + b\right), \quad \forall i = 1, \dots, m. \tag{18}$$

It also follows from the dual of the soft-margin SVM formulation (see [6]) that, for a given sample $i = 1, .., m$, the value of its Lagrange multiplier α_i is upper-bounded by the trade-off parameter C. However, if we redefine the SVM objective function from $C\Sigma_i \xi_i$ to $\frac{C}{\overline{W}} \cdot \sum_i W_i \xi_i$, then $\alpha_i \leq \frac{C \cdot W_i}{\overline{W}}$. Algorithm (2) presents the Kernel-based OWA–SVM method.

Algorithm 2 OWA-SVM Algorithm, Kernel-based version.

```
Input: Training tuples $(\mathbf{x}_i, y_i)$, SVM soft-margin parameter $C$, OWA quantifier parameter $\alpha$.
Output: SVM classifier $(\boldsymbol{\alpha}, b)$.

1. $\boldsymbol{\xi}_1^{\ddagger} \leftarrow$ Kernel-based SVM training, Formulation (15), with parameter $C$ and using Eq. (18).
2. $F(\boldsymbol{\xi}_1^{\ddagger}, \mathbf{W}) \leftarrow$ OWA operator, order induced by $\boldsymbol{\xi}_1^{\ddagger}$, and weights $\mathbf{W}$ obtained via an OWA quantifier based on quantifier parameter $\alpha$.
3. $(\boldsymbol{\alpha}, b) \leftarrow$ OWA-SVM training:

   $\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,s=1}^{m} \alpha_i \alpha_s y_i y_s K(\mathbf{x}_i, \mathbf{x}_s)$
   $\text{s.t.} \sum_{i=1}^{m} \alpha_i y_i = 0,$        (19)
   $0 \leq \alpha_i \leq \frac{C \cdot W_i}{\overline{W}}, \quad i = 1, \dots, m.$
```

Next, we studied the performance of the proposed method empirically, comparing it with well-known data mining methods on benchmark datasets from the UCI Repository.

4. Experimental results

We applied the proposed approaches to the following datasets from the UCI Repository [2]: Ionosphere (IONO), Wisconsin Breast Cancer (WBC), Australian Credit (AUS), Pima Indians Diabetes (DIA), German Credit (GC), and Splice (SPL). Table 1 summarizes the metadata for each dataset, including the total sample size and per-class proportion, and the number of features.

Table 1
Descriptive information for all datasets.

Dataset	No. of examples	% Class(min.,maj.)	No. of features
IONO	351	(64.1,35.9)	34
WBC	569	(62.7,37.3)	30
AUS	690	(55.5,44.5)	14
DIA	768	(65.1,34.9)	8
GC	1000	(70.0,30.0)	24
SPL	1000	(51.7,48.3)	60

Table 2
Performance for all methods and datasets.

	Linear/traditional methods					Kernel methods	
	k-NN	NB	Logit	SVM$_l$	OWA–SVM$_l$	SVM$_k$	OWA–SVM$_k$
IONO	79.2	83.2	83.6	84.5	**86.0**	**95.1**	94.0
WBC	95.9	92.7	95.2	97.4	**97.6**	97.7	**97.9**
AUS	84.5	79.3	85.8	86.2	**87.1**	86.5	**87.6**
DIA	70.4	71.8	72.5	72.7	**73.5**	72.3	**73.3**
GC	62.9	69.4	70.0	69.8	**70.2**	69.7	**71.1**
SPL	71.1	**84.2**	80.1	80.9	81.7	88.1	**88.5**

The following binary classification approaches are studied and reported:

- Soft-margin SVM, linear (SVM$_l$, Formulation (14)) and Kernel-based version (SVM$_K$, Formulation (15)). This method was implemented using the LIBSVM toolbox [8].
- The k nearest neighbors method (k-NN). The idea behind this approach is to assign a label to a new sample based on the labels of the k observations from the training set that are closest to this new sample, using majority voting. The Euclidean distance is used to determine the nearest neighbors [17].
- The naïve Bayes method (NB). The idea is to apply the Bayes theorem assuming that all features are independent of each other. The method computes the *a posteriori* probability for each of the two classes under this assumption, and the class with maximum probability is assigned to the new sample [17].
- The logistic regression approach (Logit). This method constructs a linear classifier, similar to linear SVM, in which the coefficients are estimated using maximum likelihood estimation [17].
- The proposed OWA-SVM method in its linear (OWA-SVM$_l$, Algorithm (1)) and Kernel-based method (OWA-SVM$_k$, Algorithm (2)).

Regarding model selection, a grid search was performed for SVM parameters C and σ (Kernel methods only), using 10-fold cross-validation with the following set of parameters: $C, \sigma \in \{2^{-7}, 2^{-6}, \dots, 2^6, 2^7\}$. For our proposal, four different quantifiers were studied: Basic RIM, quadratic, exponential, and trigonometric. We explored the following values for the quantifier parameter: $\alpha \in \{0.2, 0.4, 0.6, 0.8\}$. Finally, $k = 5$ was used for the k-NN method.

Next, a summary of the results is presented in Table 2. This table shows the performance of each method for all six datasets using AUC (×100) as the performance metric. The highest AUC is highlighted in bold type for each dataset, in which Kernel methods were studied as a different group.

In Table 2, it can be observed that the best overall performance is achieved with the proposed methods for both linear/traditional and Kernel-based approaches. For the first group, OWA-SVM$_l$ performs best in five of the six cases, being second in terms of performance on the Splice dataset. For the Kernel-based group, OWA-

Table 3
Holm's post-hoc test for pairwise comparisons. All methods.

Method	Mean Rank	Mean AUC	p value	$\alpha/(k-i)$	Action
Linear/traditional methods					
OWA-SVM$_l$	1.17	82.68	–	–	not reject
SVM$_l$	2.33	81.92	0.2012	0.0500	not reject
Logit	3.17	81.20	0.0285	0.0250	not reject
NB	3.83	80.10	0.0035	0.0167	reject
k-NN	4.50	77.33	0.0003	0.0125	reject
Kernel methods					
OWA-SVM$_k$	1.17	85.40	–	–	not reject
SVM$_k$	1.83	84.90	0.1025	0.0500	not reject

Table 4
Holm's post-hoc test for pairwise comparisons: OWA quantifiers.

Method	Mean Rank	Mean AUC	p value	$\alpha/(k-i)$	Action
OWA–SVM$_l$					
Basic Rim	1.33	83.37	–	–	not reject
Exponential	2.50	83.03	0.1175	0.0500	not reject
Quadratic	2.58	83.10	0.0935	0.0250	not reject
Trigonometric	3.58	82.78	0.0025	0.0167	not reject
OWA–SVM$_k$					
Trigonometric	2.08	85.12	–	–	not reject
Quadratic	2.17	85.03	0.9110	0.0500	not reject
Basic Rim	2.50	84.92	0.5762	0.0250	not reject
Exponential	3.25	84.85	0.1175	0.0167	not reject

Table 5
Holm's post-hoc test for pairwise comparisons: OWA quantifier parameter α.

Method	Mean Rank	Mean AUC	p value	$\alpha/(k-i)$	Action
OWA–SVM$_l$					
$\alpha = 0.2$	2.17	83.65	–	–	not reject
$\alpha = 0.4$	2.33	83.65	0.8231	0.0500	not reject
$\alpha = 0.8$	2.50	83.57	0.6547	0.0250	not reject
$\alpha = 0.6$	3.00	83.58	0.2636	0.0167	not reject
OWA–SVM$_k$					
$\alpha = 0.4$	2.25	85.10	–	–	not reject
$\alpha = 0.8$	2.25	85.08	1.0000	0.0500	not reject
$\alpha = 0.2$	2.75	85.08	0.5023	0.0250	not reject
$\alpha = 0.6$	2.75	84.93	0.5023	0.0167	not reject

Table 6
Average running times, in seconds. All methods and datasets.

	OWA–SVM$_l$		OWA–SVM$_k$		k-NN	Logit	NB
	Step 1[a]	Step 2[b]	Step 1[a]	Step 2[b]			
AUS	0″.05	4″.27	0″.10	2″.28	0″.06	0″.02	0″.01
WBC	0″.03	2″.54	0″.10	1″.69	0″.06	0″.25	0″.02
DIA	0″.04	5″.98	0″.11	3″.31	0″.02	0″.003	0″.002
GC	0″.15	6″.78	0″.26	4″.65	0″.13	0″.03	0″.03
IONO	0″.01	0″.78	0″.03	0″.57	0″.03	0″.17	0″.01
SPL	0″.24	6″.29	0″.48	3″.93	0″.20	0″.10	0″.14

[a] The first step for OWA-SVM is equivalent to soft-margin SVM implemented in LIBSVM since the computation of the OWA operator takes less than 0.01 seconds.

[b] The second step for OWA-SVM is equivalent to soft-margin SVM implemented in MATLAB's 'quadprog' solver since the computation of the OWA operator takes less than 0.01 seconds.

SVM$_k$ also performed better than SVM$_k$ in five of the six datasets; the latter performed better on the Ionosphere dataset.

Next, The Holm's test was used to study the statistical significance of the above results. This test was suggested by Demšar [11] for comparing various machine learning methods in terms of performance. The average rank is computed for each technique. Then, pairwise comparisons are performed between each method and the top-ranked one. From these comparisons, Z statistics are obtained, with their corresponding p values. Each p_i is compared with $\alpha/(k-i)$, where k is the number of algorithms and α the significance level (usually 5%, see [11] for more details). This analysis is presented in Table 3 for both groups.

From the Holm's test results reported in Table 3, it can be concluded that the proposed OWA-SVM$_l$ achieves the best overall performance, being able to outperform Naïve Bayes and k-NN statistically, with $\alpha = 5\%$. Although OWA-SVM$_l$ has an average rank of 1.17, being best in five of six datasets, it is not able to outperform either logistic regression or SVM$_l$. Similarly, OWA-SVM$_k$ cannot outperform SVM$_k$ with an average rank of 1.17.

Next, the Holm's test was used for comparing the various quantifiers studied for OWA–SVM$_l$ and OWA–SVM$_k$. We recall that four different quantifiers (Basic RIM, quadratic, exponential, and trigonometric) and four different values for α ($\alpha \in 0.2, 0.4, 0.6, 0.8$) were studied for each method. Table 4 presents the results of the Holm's test for pairwise comparisons between the quantifiers for OWA–SVM$_l$ and OWA–SVM$_k$, while Table 5 presents the same exercise for the four α values.

In Tables 4 and 5, it can be seen that no quantifier is able to outperform the others with statistical significance, all being relatively similar in terms of the average rank and AUC. It can be noticed in Table 4 that the worst quantifier for OWA–SVM$_l$ in terms of average rank is the best one for OWA–SVM$_k$ (Trigonometric quantifier), and that quantifiers are very close in terms of average AUC. It can be concluded that all quantifiers are almost equally good at identifying the right weights $\mathbf{W}$ for the OWA operator.

Finally, the running times are presented in Table 6 for all methods and datasets. For the proposed OWA–SVM, we split the running time in two steps: Step 1 corresponds to the first SVM training performed using LIBSVM, while Step 2 is the second SVM training performed using a generic QP solver (MATLAB's 'quadprog' function). Since the computation of the OWA function, which corresponds to evaluating the quantifier and sorting the samples based on the slack variables, is extremely fast (less than 0.01 seconds), the running times for OWA–SVM are equivalent to those for SVM, either using a highly-optimized solver (LIBSVM, Step 1), or a generic one ('quadprog', Step 2). Notice that these two problems have a similar complexity. These experiments were performed on an HP Envy dv6 with 16 GB RAM, 750 GB SSD, a i7-2620M processor with 2.70 GHz, and using Microsoft Windows 8.1 Operating System (64-bits). Each entry in Table 6 represents the average training time (in seconds) of all runs of the ten-fold crossvalidation procedure.

It can be observed on Table 6 that all training times are tractable and relatively similar. The only exception is the second step of our algorithm, which is considerably slow in comparison with the remaining methods and, in particular, with SVM using LIBSVM (OWA–SVM Step 1). As mentioned before, this is exclusively because of the optimization approach used since both steps are of the same complexity. The difference relies in the hyperparameter C, which is a scalar in the first step and a vector in the second one ($\frac{C \cdot W_i}{W}$, see Formulation (19)). Unfortunately, a vector is not a valid input parameter for LIBSVM, and therefore we used a generic QP solver for the second step.

5. Conclusions and future developments

In this work, a novel binary classification approach based on the concept of OWA operators was presented. The main idea is to penalize the classification errors unevenly with a weighted sum according to their distances from their respective canonical hyperplanes. An OWA operator is then constructed to replace the hinge loss function in the soft-margin SVM formulation. Various OWA quantifiers were explored empirically, showing that our proposal outperforms other well-known classification methods.

Our method is presented first as a linear classification approach, and subsequently extended as a Kernel method. Linear methods provide a better understanding of the process that generates the data, but they are not able to capture more complex patterns [17]. In our experiments we observe that the results are relatively similar among all the models studied in four of the six datasets. For the remaining two, however, Kernel methods are able to outperform all the other approaches, demonstrating the importance of Kernel methods when predictive performance is of prime interest.

There are various opportunities for future research. In terms of the OWA quantifiers, no significant differences were observed among the four quantifiers studied, according to our experiments. The same holds for the different values for the quantifier parameter α. A meta-learning study is suggested as a future development, aiming at understanding the behaviour of various quantifiers using datasets various characteristics. Another possible research line is the use of different OWA operators to induce order in the training samples. Induced OWA (IOWA), for example, allows the use of a different criterion to sort the samples, avoiding the need for the first SVM training, and thus speeding up the algorithm. Another interesting approach is the use of the weighted OWA (WOWA) operator [23,35] since it integrates the weighted average and the OWA operator in the same formulation. Finally, the proposed strategies can be used in applied contexts, such as business analytics and computer vision. Since our proposal leads to best predictive performance compared to other well-known machine learning methods, it becomes an interesting alternative in applications for which an increase in performance translates in profitable decision-making, such as credit scoring or churn prediction. The flexibility that SVM provides allows the use of profit metrics in combination with our proposal for a goal-oriented machine learning framework [20,33].

Acknowledgements

This research was partially funded by CONICYT, FONDECYT projects 1160286 and 1160738, and by the Complex Engineering Systems Institute (CONICYT, PIA, FB0816). We would also like to thank the anonymous reviewers for valuable comments that have improved the quality of the paper.

References

[1] N. Alajlan, Y. Bazi, H.S. AlHichri, F. Melgani, R.R. Yager, Using OWA fusion operators for the classification of hyperspectral images, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6 (2) (2013) 602–614.
[2] A. Asuncion, D. Newman, UCI machine learning repository, 2007, http://archive.ics.uci.edu/ml/.
[3] B.D. Barkana, I. Saricicek, B. Yildirim, Performance analysis of descriptive statistical features in retinal vessel segmentation via fuzzy logic, ann, svm, and classifier fusion, Knowl.Based Syst. 118 (2017) 165–176.
[4] G. Beliakov, A. Pradera, T. Calvo, Aggregation functions: A guide for practitioners, 221, Springer, 2007.
[5] F. Blanco-Mesa, J.M. Merigó, J. Kacprzyk, Bonferroni means with distance measures and the adequacy coefficient in entrepreneurial group theory, Knowl.Based Syst. 111 (2016) 217–227.
[6] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discovery 2 (2) (1998) 121–167.
[7] E. Byvatov, G. Schneider, Support vector machine applications in bioinformatics., Appl. Bioinf. 2 (2) (2003) 67–77.
[8] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[9] W. Chen, H. Yang, P. Feng, H. Ding, H. Lin, IDNA4mc: identifying DNA n4-methylcytosine sites based on nucleotide chemical properties, Bioinformatics 33 (22) (2017) 3518–3523.
[10] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[11] J. Demšar, Statistical comparisons of classifiers over multiple data set, J. Mach. Learn. Res. 7 (2006) 1–30.
[12] A. Emrouznejad, M. Marra, Ordered weighted averaging operators 1988–2014: a citation-based literature survey, Int. J. Intell. Syst. 29 (11) (2014) 994–1014.
[13] J. Fodor, J.-L. Marichal, M. Roubens, Characterization of the ordered weighted averaging operators, IEEE Trans. Fuzzy Syst. 3 (2) (1995) 236–240.
[14] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139.
[15] J. Gao, M. Li, H. Liu, Generalized ordered weighted utility averaging-hyperbolic absolute risk aversion operators and their applications to group decision-making, Eur. J. Oper. Res. 243 (1) (2015) 258–270.
[16] M. Grabisch, J.-L. Marichal, R. Mesiar, E. Pap, Aggregation functions: means, Inf. Sci. 181 (1) (2011) 1–22.
[17] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Technique, 3, Morgan Kaufmann, Waltham, MA, USA, 2011.
[18] M. Li, Y.-T. Zheng, S.-X. Lin, Y.-D. Zhang, T.-S. Chua, Multimedia evidence fusion for video concept detection via OWA operator, in: International Conference on Multimedia Modeling, Springer, 2009, pp. 208–216.
[19] P. Luuka, O. Kurama, Similarity classifier with ordered weighted averaging operators, Expert Syst. Appl. 40 (4) (2013) 995–1002.
[20] S. Maldonado, A. Flores, T. Verbraken, B. Baesens, R. Weber, Profit-based feature selection using support vector machines - general framework and an application for customer churn prediction, Appl. Soft Comput. 35 (2015) 740–748.
[21] S. Maldonado, J. López, Synchronized feature selection for support vector machines with twin hyperplanes, Knowl. Based Syst. 132C (2017) 119–128.
[22] S. Maldonado, R. Weber, J. Basak, Kernel-penalized SVM for feature selection, Inf. Sci. (Ny) 181 (1) (2011) 115–128.
[23] J.M. Merigó, M. Casanovas, Decision-making with distance measures and induced aggregation operators, Comput. Ind. Eng. 60 (1) (2011) 66–76.
[24] J.M. Merigó, A.M. Gil-Lafuente, The induced generalized OWA operator, Inf. Sci. (Ny) 179 (6) (2009) 729–741.
[25] J.M. Merigó, R.R. Yager, Generalized moving averages, distance measures and OWA operators, Int. J. Uncertainty Fuzziness Knowl. Based Syst. 21 (04) (2013) 533–559.
[26] J. Platt, Advances in Kernel Methods-Support Vector Learning, MIT Press, Cambridge, MA, pp. 185–208.
[27] R. Ren, ANN Vs. SVM: which one performs better in classification of MCCs in mammogram imaging, Knowl. Based Syst. 26 (2012) 144–153.
[28] R. Ribeiro, R. Alberto, N. Pereira, Generalized mixture operators using weighting functions: a comparative study with WA and OWA, Eur. J. Oper. Res. 145 (2) (2003) 329–342.
[29] B. Schölkopf, A.J. Smola., Learning with Kernels, MIT Press, 2002.
[30] C.E. Shannon, A mathematical theory of communication, ACM SIGMOBILE Mobile Comput. Commun. Rev. 5 (1) (2001) 3–55.
[31] V. Vapnik, Statistical Learning Theory, John Wiley and Sons, 1998.
[32] V. Vapnik, A. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (2) (1971) 264–280.
[33] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, B. Baesens, New insights into churn prediction in the telecommunication sector: a profit driven data mining approach, Eur. J. Oper. Res. 218 (1) (2012) 211–229.
[34] Z. Xu, J. Chen, Ordered weighted distance measure, J. Syst. Sci. Syst. Eng. 17 (4) (2008) 432–445.
[35] Z.S. Xu, Q.-L. Da, The uncertain OWA operator, Int. J. Intell. Syst. 17 (6) (2002) 569–575.
[36] R.R. Yager, On ordered weighted averaging aggregation operators in multicriteria decision making, IEEE Trans. Syst. Man Cybern. 18 (1) (1988) 183–190.
[37] R.R. Yager, Families of OWA operators, Fuzzy Sets Syst. 59 (2) (1993) 125–148.
[38] R.R. Yager, On generalized bonferroni mean operators for multi-criteria aggregation, Int. J. Approximate Reasoning 50 (8) (2009) 1279–1286.
[39] R.R. Yager, D.P. Filev, Induced ordered weighted averaging operators, IEEE Trans. Syst. Man Cybern. Part B (Cybern.) 29 (2) (1999) 141–150.
[40] R.R. Yager, J. Kacprzyk, G. Beliakov, Recent Developments in the Ordered Weighted Averaging Operators: Theory and Practice, 265, Springer Science & Business Media, 2011.
[41] J. Yang, H. Peng, Z. Pei, Filtering e-mail based on fuzzy support vector machines and aggregation operator, in: Neural Information Processing, Springer, 2006, pp. 882–891.
[42] D. Yu, A scientometrics review on aggregation operator research, Scientometrics 105 (1) (2015) 115–133.
[43] S. Zeng, J.M. Merigó, W. Su, The uncertain probabilistic OWA distance operator and its application in group decision making, Appl. Math. Model. 37 (9) (2013) 6266–6275.
[44] S. Zeng, W. Su, A. Le, Fuzzy generalized ordered weighted averaging distance operator and its application to decision making, Int. J. Fuzzy Syst. 14 (3) (2012) 402–412.
[45] W. Zhang, T. Yoshida, X. Tang, Text classification based on multi-word with support vector machine, Knowl. Based Syst. 21 (8) (2008) 879–886.