Support Vector Machine-based approach for multi-labelers problems

S. Murillo[1] * D.H. Peluffo[1] G. Castellanos [1]

1- Universidad Nacional de Colombia -
Departamento de Ingeniería Eléctrica, Electrónica y Computación
Manizales - Colombia

Abstract. We propose a first approach to quantify the panelist's labeling generalizing a soft-margin support vector machine classifier to multi-labeler analysis. Our approach consists of formulating a quadratic optimization problem instead of using a heuristic search algorithm. We determine penalty factors for each panelist by incorporating a linear combination in the primal formulation. Solution is obtained on a dual formulation using quadratic programming. For experiments, the well-known Iris with multiple simulated artificial labels and a multi-label speech database are employed. Obtained penalty factors are compared with both standard supervised and non-supervised measurements. Promising results show that proposed method is able to asses the concordance among panelists considering the structure of data.

1 Introduction

In several supervised pattern recognition problems, a ground truth is beforehand known to carry out a training process. Nonetheless, there are cases where such ground truth is not unique. For instance, in medical environments, the diagnostic judgment given by only one doctor (panelist) might not be enough since the labeling is greatly related to the panelist's sensitivity and criterion [1]. In particular, to assess the voice quality, the panelist typically do the labeling according to their hearing abilities, ant this is certainly a subjective aspect which may complicate the design of a pattern recognition system. Few works have been concerned about this issue. In [2], authors consider a set of experts to determine the crater distribution in venus surface, by comparing human and algorithmic performance as opposed to simply comparing humans to each other. Moreover, the multi-labeler approach is only the average of labels. Other studies, [3], are focused on building proper decision boundaries from multiple-experts labels, but requiring some prior information. The approaches proposed in [4] take into account a public labeling from web pages, then the panelist confidence is not guaranteed. Finally, in [5], the multi-expert task is addressed by a support vector machine (SVM) scheme yielding a suitable approach to measure panelist performance.

This work proposes a first methodology to quantify the panelist's labeling from a soft-margin support vector machine approach (SMSVM), as a variation to that proposed in [5]. Such variation consists of formulating the optimization problem within a quadratic programming framework instead of using a heuristic search algorithm, as usual. Our method's outcomes are penalty or relevance values associated to each panelist, pointing out a well performing labeler when lower is its value. For experiments,

*This work is supported by the "Aprendizaje de máquina a partir de múltiples expertos en clasificación multiclase de señales de voz" project associated with "Jóvenes Investigadores" program by COLCIENCIAS

two databases are considered. Firstly, the well-known Iris with multiple artificial labels. Secondly, a multi-labeler speech database for detecting hypernasality. Obtained penalty factors are compared with both standard supervised and non-supervised measurements. The results are promising being our method able to asses the concordance among panelists taking into account the structure of data. This paper is organized as follows: In section 2, we briefly describe our method to analyze the reliability of panelist labeling. Section 3 shows and discuss the obtained results. Finally, in section 4, some final remarks and conclusions are presented.

2 Multi-labeling analysis based on a binary SVM formulation

Our approach consists of a variation of a SVM two-class (binary) classifier and works as follows: We start assuming a latent variable model, which is to be used as the classifier decision function. Then, we formulate an optimization problem by generalizing the classifier taking into account different labeling vectors and adding penalty factors in a similar frameworks as that described in [5]. Define the ordered pair $\{\boldsymbol{x}_i, y_i\}$ to represent the i-th sample where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the d-dimensional feature vector and y_i is the binary corresponding class label for two classes problem, such that $y_i \in \{1, -1\}$. In matrix terms, $\boldsymbol{X} \in \mathbb{R}^{m\times d}$ and $\boldsymbol{y} \in \mathbb{R}^m$, are respectively the data matrix and labeling vector, being d the number of considered features and m the number of samples. We assume an hyperplane model in the form: $\boldsymbol{w}\cdot\boldsymbol{x}+b = \boldsymbol{w}^\top\boldsymbol{x}+b = 0$, where $\boldsymbol{w}$ is an orthogonal vector to the hyperplane, b is a bias term and notation $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product. Intuitively, for a two-class problem we can establish $f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^\top\boldsymbol{x}+b)$ as a decision function. In order to avoid that data points lie in a region where there exists ambiguity to take the decision, we assure that the distance between the hyperplane and any data point to be at least 1, satisfying the condition: $y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)) \geq 1 \quad \forall i$. Then, the distance between any data point x_i and the hyperplane $(\boldsymbol{w}, b)$ can be calculated as: $\text{d}\left((\boldsymbol{w}, b), \boldsymbol{x}_i\right) = y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)/||\boldsymbol{w}||_1^2 \geq 1/||\boldsymbol{w}||_2^2$, where $||\cdot||_2$ stands for Euclidean norm. Therefore, we expect that $y_i \simeq \boldsymbol{w}^\top\boldsymbol{x}_i+b$, since upper boundary is $1/||\boldsymbol{w}||^2$. Then, the classifier objective function to be maximized can be written as: $\max_{\boldsymbol{w}} y_i(\boldsymbol{w}^\top\boldsymbol{x_i}+b)/||\boldsymbol{w}||^2; \quad \forall i$. For accounts of minimization, we can re-write the above problem so:

$$\min_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w}||^2, \quad \text{s.t.} \quad y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b) = 1; \ \ \forall i \tag{1}$$

By relaxing (1), we can write the following SVM-based formulation:

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}|\lambda, b) = \min_{\boldsymbol{w}} \frac{\lambda}{2}||\boldsymbol{w}||^2 + \frac{1}{m}\sum_{i=1}^{m}(1 - \boldsymbol{y}_i(\boldsymbol{w}^\top\boldsymbol{x}+b))^2 \tag{2}$$

where λ is a regularization parameter.

2.1 Soft margin

Previous formulation is a hard margin approach, i.e., data points are not expected to lie on the decision function boundary. Recalling (2), we can extend the functional to a soft

margin formulation incorporating a slack variable ξ_i, such that: $1 - y_i < \boldsymbol{x_i}, \boldsymbol{w} > \leq \xi_i;\ \forall i$. Note that in this approach, we assume $b = 0$. In accordance with this framework, we can write a soft margin SVM formulation (SMSVM) as:

$$\min_{\boldsymbol{w},\boldsymbol{\xi}} \hat{f}(\boldsymbol{w}, \boldsymbol{\xi}|\lambda) = \min_{\boldsymbol{w},\boldsymbol{\xi}} \frac{\lambda}{2}||\boldsymbol{w}||^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i^2, \quad \text{s.t.} \quad \xi_i \geq 1 - y_i\langle \boldsymbol{x}_i, \boldsymbol{w}\rangle \tag{3}$$

where $\boldsymbol{\xi} \in \mathbb{R}^m = [\xi_1, \ldots, \xi_m]$.

Since in problem stated in (2), term $1 - y_i < \boldsymbol{x}_i, \boldsymbol{w} >$ has the upper boundary at ξ_i, minimizing $f(\cdot)$ regarding $\boldsymbol{w}$, is the same as minimizing $\hat{f}(\cdot)$ with respect to $\boldsymbol{w}$ and ξ_i.

2.2 Multi-labeler analysis

To address the matter that we are concerned about in this work, we aim to design a suitable supervised classifier from the information given by different sources (labeling vectors). In this work, we propose to incorporate a penalty factor θ_t, such that $\hat{f}(\cdot)$ decreases when adding right labels otherwise it should decrease. This approach is done in a similar way as that proposed in [5] but using a quadratic version. Consider a set of k panelists who assign their corresponding labeling vectors. Then, the t-th panelist is to be associated to penalty factor θ_t, where $t \in [k]$ and $[k] = \{1, \ldots, k\}$. Accordingly, by including the penalty factor $\boldsymbol{\theta}$, we can re-write the functional given in 2 as:

$$\min_{\boldsymbol{w},\boldsymbol{\xi}} \frac{\lambda}{2}||w||_2^2 + \frac{1}{2m}\sum_{i=1}^{m}(\xi_i + \frac{1}{k}\sum_{t=1}^{k} c_{ij}\theta_t)^2, \ \text{s.t.} \quad \xi_i \geq 1 - y_i\langle \boldsymbol{x}_i, \boldsymbol{w}\rangle - \frac{1}{k}\sum_{t=1}^{k} c_{it}\theta_t \tag{4}$$

where c_{it} is the coefficient for the linear combination of all θ_t representing the relevance of the information given by t-th panelist over the sample i, defined as:

$$c_{it} = \frac{n_e(\boldsymbol{y}^{(t)} = \boldsymbol{y}_{ref})}{m}|\boldsymbol{w}^\top \boldsymbol{x}_i|. \tag{5}$$

Defining an auxiliary variable $\hat{\xi}_i$ as $\hat{\xi}_i = \xi_i + \frac{1}{k}\sum_{t=1}^{k} c_{it}\theta_t \Rightarrow \hat{\boldsymbol{\xi}} = \boldsymbol{\xi} + \frac{1}{k}\boldsymbol{C}\boldsymbol{\theta}$,

$$\min \frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w} + \frac{1}{2m}\hat{\boldsymbol{\xi}}^\top \hat{\boldsymbol{\xi}}, \quad \text{s.t.} \quad \hat{\boldsymbol{\xi}} \geq \boldsymbol{1}_m - (\boldsymbol{X}\boldsymbol{w}) \circ \boldsymbol{y} \tag{6}$$

Assuming the critical case $\hat{\boldsymbol{\xi}} = \boldsymbol{1}_m - (\boldsymbol{X}\boldsymbol{w}) \circ \boldsymbol{y}$, the corresponding Lagrangian of (4) is:

$$\mathcal{L}(\boldsymbol{w}, \hat{\boldsymbol{\xi}}|\lambda) = \hat{f}(\boldsymbol{w}, \boldsymbol{\xi}|\lambda) + g(\hat{\boldsymbol{\xi}}, \boldsymbol{w})^\top \boldsymbol{\alpha} = \frac{\lambda}{2}\boldsymbol{w}\boldsymbol{w}^\top + \frac{1}{2m}\boldsymbol{\xi}^\top \boldsymbol{\xi} + (\boldsymbol{\xi} - \boldsymbol{1}_m + (\boldsymbol{X}\boldsymbol{w}) \circ \boldsymbol{y})^T \boldsymbol{\alpha}.$$

Now, solving the Karush-Kuhn-Tucker conditions, we have:
$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \lambda \boldsymbol{w} + (\boldsymbol{x}^\top \circ (\boldsymbol{y}^\top \otimes \boldsymbol{1}_d))\boldsymbol{\alpha} = 0 \Rightarrow \boldsymbol{w} = -\frac{1}{\lambda}(\boldsymbol{x}^\top \circ (\boldsymbol{y}^\top \otimes \boldsymbol{1}_d))\boldsymbol{\alpha}$,
$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}} = \frac{1}{m}\hat{\boldsymbol{\xi}} + \boldsymbol{\alpha} = 0 \Rightarrow \hat{\boldsymbol{\xi}} = -m\boldsymbol{\alpha}$,
$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}} = \hat{\boldsymbol{\xi}} - \boldsymbol{1}m + (\boldsymbol{X}\boldsymbol{w}) \circ \boldsymbol{y} = 0$,
where $\boldsymbol{\alpha}$ is the vector of lagrange multipliers. Under these conditions and eliminating the primal variables from (6), we can pose a new problem in terms of the dual variable $\boldsymbol{\alpha}$, so:

$$\min_{\boldsymbol{\alpha}} \hat{f}(\boldsymbol{\alpha}|\lambda) = \frac{1}{2\lambda}\boldsymbol{\alpha}^\top \boldsymbol{P}\boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \boldsymbol{D}\boldsymbol{\alpha} - \mathbf{1}_m^\top, \quad \text{s.t.} \quad \boldsymbol{\alpha} > 0 \tag{7}$$

where

$$\boldsymbol{P} = (\boldsymbol{x}^\top \circ (\mathbf{1}_d \otimes \boldsymbol{y}^\top))^\top ((\boldsymbol{x}^\top \circ (\mathbf{1}_d \otimes \boldsymbol{y}^\top))), \; \boldsymbol{D} = (-\frac{1}{\lambda}(\boldsymbol{x} \circ (\mathbf{1}_d \otimes \boldsymbol{y}^\top))^\top) \circ ((\mathbf{1}_d \otimes \boldsymbol{y}^\top) \circ \boldsymbol{x}^\top)$$

As it can be appreciated, formulation given by (7) is an evident quadratic problem with linear constraints, which can be solved by means of a heuristic usually applied for quadratic programming methods. Finally, θ value is calculated by this way:

$$\boldsymbol{\theta} = \boldsymbol{C}^\dagger(1 - \boldsymbol{y} \circ (\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{\xi}) \tag{8}$$

3 Results and discussion

For experiments, two databases are used. Firstly, Iris database from UCI repository, for which the first 100 samples being linear separable clases are taken into account. As $\boldsymbol{y}_{\text{ref}}$, the original labels are considered. Additionally, 7 simulated labeling vectors are built to represent different panelists. Secondly, the Hypernasality database provided by Control and Digital Signal Processing research group from Universidad Nacional de Colombia – Manizales. This database is formed by 156 samples from children pronouncing in Spanish the vowel $/a/$. Samples are characterized as detailed in [6] in order to obtain the feauture space to be analyzed. Labels are made by a speech terapist team conformed by 3 experts. In this case, since there is no a reference labeling vector, it is estimated as: $\boldsymbol{y}_{\text{ref}} = \text{sign}(1/k \sum_{t=1}^{k} \boldsymbol{y}^{(t)})$. For comparison purposes, we employ standard supervised measurements (sensitivity (Se), specificity (Sp) and classification performance (CP)). As well, unsupervised measurements Fisher's criterion (J) and silhouette (S). To solve the quadratic formulation, an active-set algorithm is employed.

Fig. 1 depicts the results obtained over the Iris database, where Fig. 1(j) show the reference labeling vector and 7 simulated panelist as well, figures between 1(a) and 1(h) show the corresponding scatter plots for all the panelists and the boundary decision given by the our method. Fig. 1(i) shows the $\boldsymbol{\theta}$ values representing the weights or penalty factors associated to each labeler, as well as the Fisher's criterion and silhouette values. In general, low values for fisher and silhouette point out that clusters are not homogeneous, meanwhile high values refers compactness. In turn, the $\boldsymbol{\theta}$ values decrease when the labeling error is lower, otherwise it increases. Then, it is possible to say that lower values of $\boldsymbol{\theta}$ imply higher values for the two considered measurements. In particular, in Fig. 1, we can appreciate that those panelists exhibiting higher error are strongly penalized. In contrast, those ones with lower error have small penalty factors.

Fig. 2 shows the performance of our method over the hypernasality database. Despite that the classes are not linearly separable, we can observe that the decision boundary obtained by proposed method seems to be a good separation.

Fig. 2 depicts the scatter plots for labelers as well as the reference labeling vector. Besides, in Fig. 2(e), the $\boldsymbol{\theta}$, J and S values for this experiment are presented. The corresponding label set is depicted Fig. 2(f). As it can be appreciated, those labelers closest

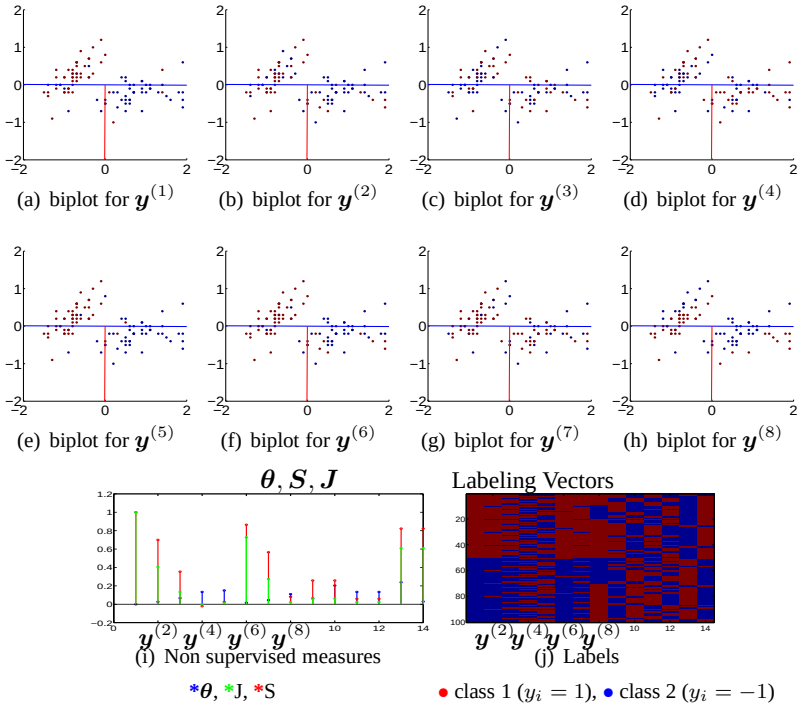

Fig. 1: Iris Database Experiment

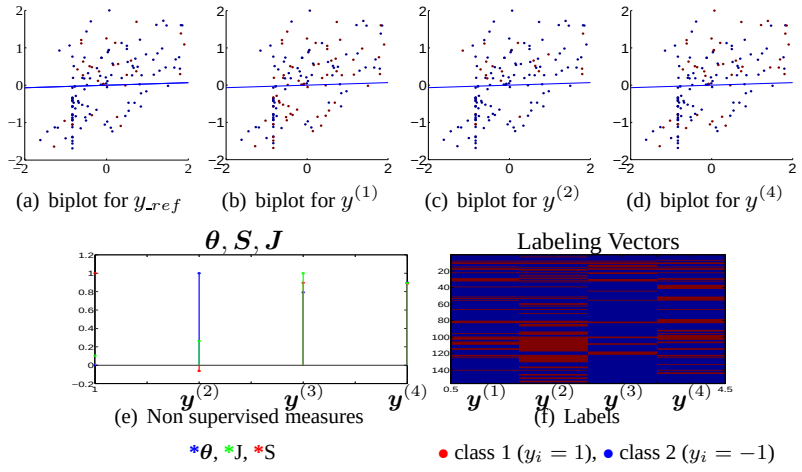

Fig. 2: Hipernasality Database Experiment

to cluster separation given by the SVM classifier are penalized with lower weights than those ones distant from it. Also, we can notice that panelist $\boldsymbol{y}^{(3)}$ achieves more compact

clusters in comparison with the remaining panelists, and this is evidenced by factor $\boldsymbol{\theta}$.

Table 1: Comparison between penalty factors and other measures

	θ_t	S	J	Se	Sp	CP
$\boldsymbol{y}_{ref}$	0	0.10	1.00	1	1	1
$\boldsymbol{y}^{(1)}$	1	0.26	-0.06	0,94	0,44	0,72
$\boldsymbol{y}^{(2)}$	0.58	1	0.89	0,86	0,69	0,83
$\boldsymbol{y}^{(3)}$	0.46	0.89	0.88	0,96	0,65	0,87

Table 1 shows some numerical results. It is important to highlight that the proposed method provides penalty factors in a such way that the distances between each data point and the hyperplane is considered. For this reason, our method's performance is sensitive to the feature space and the selection of reference labeling vector.

4 Conclusions and future work

Experimentally, we proved that the proposed approach is capable to quantify the confidence of a set of panelist taking into consideration the natural structure of data. This is done by penalizing the supposed wrong labels regarding the distance between its corresponding data point and a decision hyperplane. This approach might allow to identify those panelists or data points supposed to be untrustful labeler as well as outliers.

For future works, we are aiming to explore alternatives to improve the reference labeling vector setting, since the simple average may not be an adequate reference for all cases, specially, when there are many supposed wrong labelers.

References

[1] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, Linda Moy, and David Blei. Learning from crowds.

[2] Padhraic Smyth, Usama M. Fayyad, Michael C. Burl, Pietro Perona, and Pierre Baldi. Inferring ground truth from subjective labelling of venus images. In *NIPS'94*, pages 1085–1092, 1994.

[3] Koby Crammer, Michael Kearns, Jennifer Wortman, and Peter Bartlett. Learning from multiple sources. In *In Advances in Neural Information Processing Systems 19*, 2007.

[4] Ofer Dekel and Ohad Shamir. Vox populi: Collecting high-quality labels from a crowd. In *In Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[5] Ofer Dekel and Ohad Shamir. Good learners for evil teachers. In *ICML*, page 30, 2009.

[6] J. R. Orozco, S. Murillo, A. Álvarez, J. D. Arias, E. Delgado, F. Vargas, and C. G. Castellanos. Automatic selection of acoustic and non-linear dynamic features in voice signals for hypernasality detection. In *INTERSPEECH*, pages 529–532, 2011.