Multi-labeler Analysis for Bi-class Problems Based on Soft-Margin Support Vector Machines

S. Murillo-Rendón[1], D. Peluffo-Ordóñez[1], J.D. Arias-Londoño[2], and C.G. Castellanos-Domínguez[1]

[1] Universidad Nacional de Colombia – Manizales, Caldas, Colombia
[2] Universidad de Antioquia, Medellín, Colombia

Abstract. This work presents an approach to quantify the quality of panelist's labeling by means of a soft-margin support vector machine formulation for a bi-class classifier, which is extended to multi-labeler analysis. This approach starts with a formulation of an objective function to determine a suitable hyperplane of decision for classification tasks. Then, this formulation is expressed in a soft-margin form by introducing some slack variables. Finally, we determine penalty factors for each panelist. To this end, a panelist's effect term is incorporated in the primal soft-margin problem. Such problem is solved by deriving a dual formulation as a quadratic programming problem. For experiments, the well-known Iris database is employed by simulating multiple artificial labels. The obtained penalty factors are compared with standard supervised measures calculated from confusion matrix. The results show that penalty factors are related to the nature of data, allowing to properly quantify the concordance among panelists.

Keywords: Bi-class classifier, multi-labeler analysis, quadratic programming, support vector machines.

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 (labeler, panelist, teacher, evaluator) might not be enough since the labeling is greatly related to the panelist's subjectivity and criterion [1]. Another example is the labeling carried out through internet web servers, where a set of labeler are asked to qualify data from different sources aiming to classify them in a determined group. However, because the evaluation is not restricted, that is to say, some non-experts labelers can also label the analyzed data, the additional problem of presence of noisy labels must be considered [2]. Some works have been concerned about this issue. In [3], 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 labels average. Other studies, are focused on building proper decision boundaries from multiple-experts labels, but requiring some prior information, [4]. Finally, in [5], the multi-experts task is addressed by a support vector machine (SVM) scheme yielding a suitable approach to

J.M. Ferrández Vicente et al. (Eds.): IWINAC 2013, Part I, LNCS 7930, pp. 274–282, 2013.

penalize panelist mistakes, this methodology gives a good classifier from a set of teachers, compensating the possible labeling errors by evaluating the amount and disposition of support vectors provided for each teacher to the decision boundary.

This work proposes a methodology to quantify the panelist's labeling from a soft-margin support vector machine approach (SMSVM), as a variation to that proposed in [5], the main difference in our optimization problem formulation is done within a quadratic programming framework for estimating penalty factors for labelers instead of a decision function for classification. Moreover, the proposed approach also allows to get a suitable classifier with multi-labeler training, which is further used to determine an estimation of ground truth. Also, it allows for obtaining penalization factors, one for each panelist. Such factors keep the relation between labelers and the structure of data, making evident the labeling vector quality for each labeler.

The outline of this paper is as follows: In section 2, we briefly describe our method to analyze the reliability of panelist labeling. Section 4 shows and discuss the obtained results. Finally, in section 5, some final remarks and conclusions are presented.

2 Methods

The present work proposes a variation of a SM-SVM traditional formulation is introduced, which consists of adding a penalty term and a quadratic constrain in the functional of primal formulation. Such penalty term is aimed to penalize the supposed wrong-labeled data, by means of a linear combination of some penalty factors quantifying the labeler's performance. With this variation, our approach also generates a suitable decision function allowing to obtain an estimation of ground truth, even when the labelers are wrong. In summary, we formulate an optimization problem by generalizing the classifier taking into account different labeling vectors and adding penalty factors.

2.1 Soft Margin Binary SVM Formulation

Let us define the ordered pair $\{\boldsymbol{x}_i, y_i\}$ as the i-th sample where $\boldsymbol{x}_i \in \mathbb{R}^d$ is a d-dimensional feature vector and y_i is its binary class indicator. In this case, $y_i \in \{1, -1\}$. 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.

Intuitively, for a two-class problem we can establish $f(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b)$ as the decision function. In order to avoid that data point lie in a region where there exists ambiguity to take the decision, we assure that the distance between the hyperplane and any data point must be equal or more than a priori fixed value (i.e,. one value is chosen) to satisfy 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: $\mathrm{d}\left((\boldsymbol{w}, b), \boldsymbol{x}_i\right) = y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)/||\boldsymbol{w}||_1^2 \geq 1/||\boldsymbol{w}||_2^2$, 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:

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

By relaxing (1) using a quadratic constrain, 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. 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: $1 - y_i < \boldsymbol{x_i}, \boldsymbol{w} > \leq \xi_i; \ \forall i$.

2.2 Multi-labeler Analysis Based on SM-SVM Formulation

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 increase. 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:

$$\min_{\boldsymbol{w},\boldsymbol{\xi}} \frac{\lambda}{2}||w||_2^2 + \frac{1}{2m}\sum_{i=1}^{m}(\xi_i + \frac{1}{k}\sum_{t=1}^{k} c_{ij}\theta_t)^2,$$

$$\text{s.t.} \quad \xi_i \geq 1 - y_i\langle \boldsymbol{x}_i, \boldsymbol{w}\rangle - \frac{1}{k}\sum_{t=1}^{k} c_{it}\theta_t$$

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. This term decreases the effect of the wrong labels in the construction of the boundary decision. So

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

Here $\mathrm{n_e}$ is the number of elements that satisfy the condition and Y_ref is the reference labeling vector, normally this parameter is the majority vote of the involved evaluators. In these terms, c_{it} weighted the difference between each evaluator and the reference labeling vector and also, it takes into account the distance of each sample and the decision hyperplane. Defining an auxiliary variable $\hat{\xi}_i$ as

$$\hat{\xi}_i = \xi_i + \frac{1}{k}\sum_{t=1}^{k} c_{it}\theta_t$$

The corresponding $\hat{\xi}_i$ vector formulation is

$$\hat{\boldsymbol{\xi}} = \boldsymbol{\xi} + \frac{1}{k}\boldsymbol{C}\boldsymbol{\theta} \tag{3}$$

Then, the problem formulation will be

$$\min \frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w} + \frac{1}{2m}\hat{\boldsymbol{\xi}}^\top \hat{\boldsymbol{\xi}}, \tag{4}$$

$$\text{s.t.} \quad \hat{\boldsymbol{\xi}} \geq \mathbf{1}_m - (\boldsymbol{X}\boldsymbol{w}) \circ \boldsymbol{y} \tag{5}$$

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

$$\mathcal{L}(\boldsymbol{w}, \hat{\boldsymbol{\xi}}|\lambda) = \frac{\lambda}{2}\boldsymbol{w}\boldsymbol{w}^\top + \frac{1}{2m}\boldsymbol{\xi}^\top \boldsymbol{\xi} + (\boldsymbol{\xi} - \mathbf{1}_m + (\boldsymbol{X}\boldsymbol{w}) \circ \boldsymbol{y})^T \boldsymbol{\alpha} \tag{6}$$

Now, solving the Karush-Kuhn-Tucker conditions, we have:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \lambda \boldsymbol{w} + (\boldsymbol{x}^\top \circ (\boldsymbol{y}^\top \otimes \mathbf{1}_d))\boldsymbol{\alpha} = 0 \Rightarrow \boldsymbol{w} = -\frac{1}{\lambda}(\boldsymbol{x}^\top \circ (\boldsymbol{y}^\top \otimes \mathbf{1}_d))\boldsymbol{\alpha},$$
$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}} = \frac{1}{m}\hat{\boldsymbol{\xi}} + \boldsymbol{\alpha} = 0 \Rightarrow \hat{\boldsymbol{\xi}} = -m\boldsymbol{\alpha}$$
$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}} = \hat{\boldsymbol{\xi}} - \mathbf{1}m + (\boldsymbol{X}\boldsymbol{w}) \circ \boldsymbol{y} = 0,$$

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:

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

with

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

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

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

As a remarkable matter, it is important to mention that factors $\boldsymbol{\theta}$ is greater as the amount of wrong labels increase, in accordance with equation (8). This implies that the labelers are strongly penalized when their labeling vectors scape from the ratio of the reference label vector, in this case, majority vote vector. Vector $\boldsymbol{\theta}$ also depends on the distance between each sample and the estimated decision boundary. In the case when the labeling vector corresponding to a specific labeler match with the reference vector, the θ value should be 0.

3 Experimental Setup

3.1 Database

For experiments, the Iris database from UCI repository [6] is used. We take the first 100 samples being the two linear separable clases, namely `setosa` and `versicolor` classes. A simulated labeling vector set is built to emulate different labelers. The reference labeling vector is estimated as:

$$\boldsymbol{y}_{\text{ref}} = \text{sign}\left(\frac{1}{k}\sum_{t=1}^{k}\boldsymbol{y}^{(t)}\right),$$

which corresponds to the majority vote.

3.2 Experiments

To test the methodology performance, four experiments are made. The difference between the experiments is the considered error percentage (ε) added to each simulated label set. In the first experiment, the simulated labels are built keeping the same ε for the two classes (balanced error). The second and third experiments use a label simulated set that only include error in a class (unbalanced error), second experiment affects a class and third experiment affects the another class. The last one, uses a combination of error where the second class is affected with the double of error porcentaje that the another class, it is the worst scenario. In all the experiments the ε is varied in the range from 5% to 45% by adding 5% per iteration (that is, a total of nine different ε). Additionally, for each specific ε are simulated 100 labelers. Then, the classifier $\boldsymbol{w}$ and a set of predicted labels are calculated, it is carried out by taking randomly a number t of labeling vectors from the 100 simulated labelers corresponding to this ε. The last procedure is repeated 50 times for each considered error. Predicted labels, allows to make confusion matrixes in relation to simulated labels, for that reason the data mean and standard deviation is calculated for the accuracy, sensibility and specificity from the confusion matrixes, it enable to measure the methodology performance.

4 Results and Discussion

Fig. 1(b) presents the original labels, namely the labels assigned in the Iris database for each class, the Fig. 1(c) give the majority vote, of the iterations with a ε of 20%, in this figure it is evident the description in 3.2 for the first experiment, both classes have the same ε. In Fig. 1(d) the corresponding estimated labels are showed, it is possible to note that the original and estimated labels are similar, it demostrates that the methodology is capable to find, the original labels from a noisy set of expert labels.

Meanwhile. Fig. 1(a) corresponds with the accuracy calculated for the first experiment, the methodology performance maintains a good value until the 35% in εs, it is a good yield because inclusive in high presence of wrong labels, the original labeling vectors are recovered. From a ε superior to 35% the standard deviation grows, but even

it keeps a good result. Values above 40% indicate an intolerable amount of wrong labels and it is a clear indicator of an inexpert evaluator. In Fig. 1(d) also can be seen that two samples have problems to be estimated, this samples correspond with samples nearby to the desicion boundary, just in this samples exists some ambiguity precisely by the nearness between the classes.

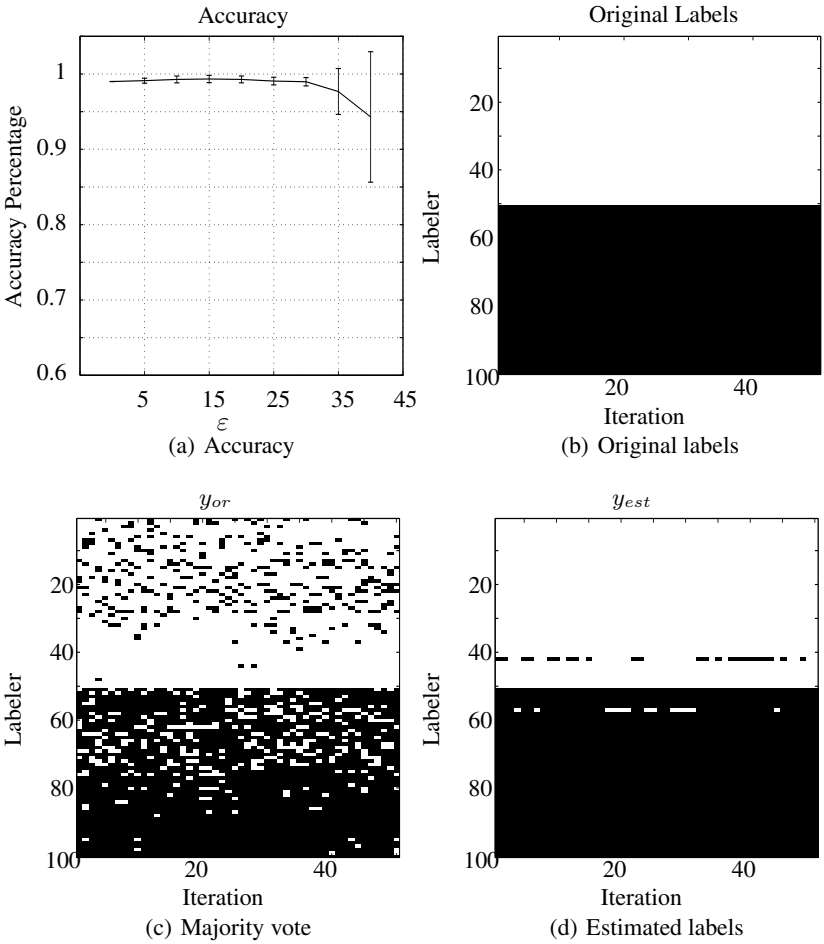

Fig. 1. Experiment1: accuracy; original, majority vote and estimated labels

In Fig. 2 it is important to observe for the experiments 2 and 3, namely, those where the noisy labels are introduced only in a class, that the estimated labels (Figs.2(g) and 2(h)) highly match with the original labels 1(b), it indicates, that when only a class is wrong labeled, it is posible to get a confiable decision boundary, that is also appreciated in table 1 and figures 2(a) and 2(b) where it is possible to find that the accuracy, sensibility and specificity are stables, it is an important result, because in several pattern recognition scenarios for automatic detection of pathologies, is common to find that

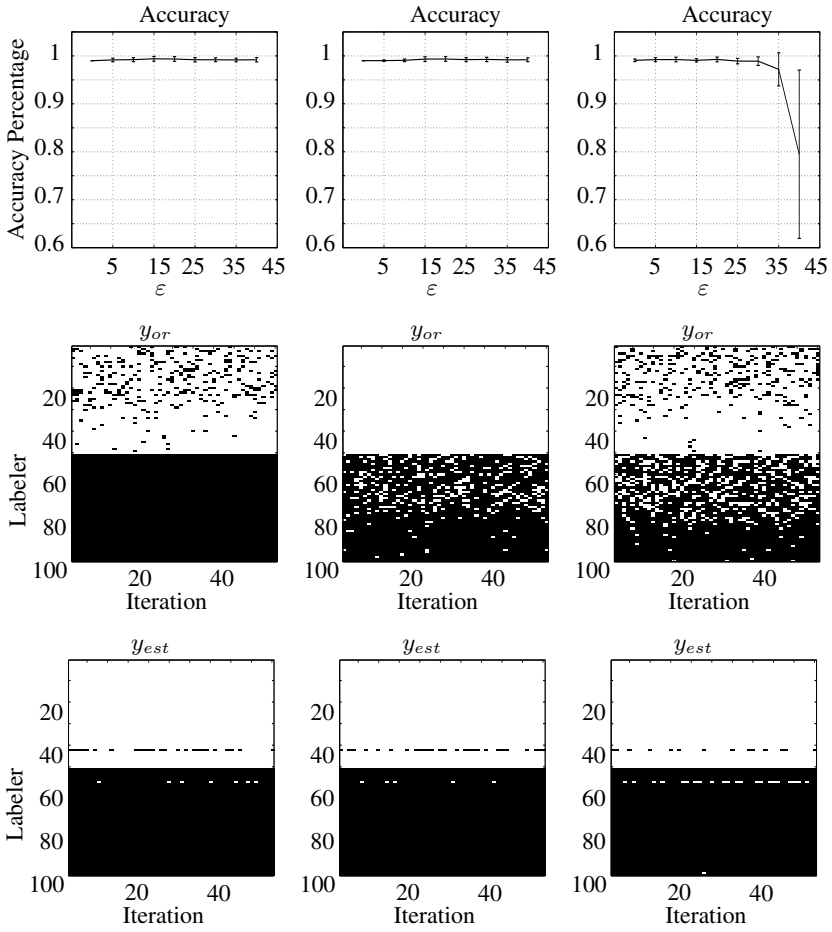

Fig. 2. Experiments Results

labelers (in many times medical experts) can make good diagnosis of normal patients, however the objective class ussually is wrong diagnosed. The present methodology can be used in these environments of automatic detection, because when only a class is wrong labeled, it is posible to build an efficient desicion boundary.

Figures 2(c), 2(f) and 2(i) shows the four experiment, it is the worst scenario for the present methodology, because there are many desbalanced noisy labels, nonetheless, can be apreciated that the estimated labels are correctly calculated inclusive for ε same 30% by a class and superior to 45% for the another class. The accuracy figure shows that only with values up to 40% the dispersion is large enough to consider a bad estimation of the desicion boundary, for this ε the accuracy value is inferior to the 80%, nevertheless, it is important to remember that this is the most difficult training environment for this methodology and with labeling error percentages as large as the exibit in this experiment, it is not possible to work in pattern recognition. Table 1 sumarize the

Table 1. Accuracy, sensibility and specificity experiments 1 and 4

Error percentage	Experiment 1			Experiment 4		
	Accuracy	Sensibility	Specificity	Accuracy	Sensibility	Especificity
0	99.0±0.00	100.0±0.00	98.0±0.00	99.1±0.27	100.0±0.00	98.2±0.55
5	99.1±0.33	100.0±0.00	98.2±0.66	99.2±0.42	100.0±0.00	98.4±0.84
10	99.3±0.45	100.0±0.28	98.6±0.93	99.2±0.48	99.7±0.74	98.8±0.99
15	99.3±0.48	99.5±0.86	99.2±1.00	99.1±0.34	99.6±0.78	98.5±0.89
20	99.3±0.45	99.5±0.89	99.1±1.01	99.3±0.49	99.1±1.01	99.4±0.91
25	99.1±0.51	99.0±1.01	99.2±1.00	**98.9±0.57**	**98.8±1.28**	**99.0±1.01**
30	99.0±0.55	98.7±1.26	99.2±0.98	**98.9±0.89**	**98.4±1.85**	**99.4±0.93**
35	97.7±3.05	96.7±3.95	98.7±2.51	**97.2±3.46**	**96.1±4.64**	**98.3±2.72**
40	94.3±8.66	92.8±9.50	95.8±8.01	**79.5±17.57**	**77.7±17.62**	**81.3±17.92**

Table 2. Accuracy, sensibility and specificity experiments 2 and 3

Error percentage	Experiment 2			Experiment 3		
	Accuracy	Sensibility	Especificity	Accuracy	Sensibility	Especificity
0	**99.0±0.00**	100.0±0.00	98.0±0.00	**99.0±0.00**	100.0±0.00	98.0±0.00
5	**99.2±0.37**	100.0±0.00	98.3±0.74	**99.0±0.14**	100.0±0.00	98.0±0.28
10	**99.2±0.42**	100.0±0.28	98.5±0.86	**99.1±0.27**	100.0±0.28	98.2±0.61
15	**99.4±0.49**	99.8±0.55	99.0±1.01	**99.3±0.48**	99.8±0.61	98.9±1.00
20	**99.4±0.48**	99.7±0.70	99.0±1.01	**99.4±0.48**	99.8±0.61	98.9±1.01
25	**99.2±0.45**	99.4±0.91	99.0±1.01	**99.2±0.42**	99.6±0.84	98.9±1.00
30	**99.2±0.40**	99.1±1.01	99.3±0.96	**99.3±0.44**	99.8±0.66	98.8±0.98
35	**99.2±0.37**	98.8±0.99	99.5±0.86	**99.2±0.48**	99.2±0.99	99.2±1.00
40	**99.2±0.45**	98.8±1.00	99.6±0.84	**99.2±0.40**	99.2±0.98	99.2±1.00

results for experiments 1 and 4, this experiments are similar in the sense of both classes are contaminated with noisy labels. In the two experiments, the desicion boundary estimation is weak for contamination above to 30% where the deviation grows quickly. Specially, the experiment 4 is affected, the bolded values show that when the evaluators have labeling inconsistences for both classes, it is difficult calculate properly the boundaries.

In table 2 the bolded values allow appreciate that for experiments 2 and 3 the accuracy, sensibility and specificity are stable along every ε, this stability is optimal, specially if it is considered that in several pattern recognition problems it is common to find a only one class good labeled. In this table also it is notable the highlighted values, because show that in the conditions of this experiments the sensibility remain in

99% up to 30% of ε, this result is important since the sensibility evaluate as good is the methodology on relation with the objective class, in automatic pathological detection terms, the sensibility measure as the methodology indentify the pathologic patients.

5 Conclusions and Future Work

Experimentally, we proved that the proposed approach is capable to retort the original labels even when the training labels contains a considerable ε level. The penalization term attached to the standard SM-SVM formulation, allows to reduce the effect of the noisy labels in the new formulation training. The experiments presented show that the present methodology is able to be used in aid systems for pathologies detection, principally because is capable of maintain good values for sensibility even when important amount of wrong labels, it favors the detection of instances in the objective class.

It is important to note that, when only a class is affected for the wrong labels the methodology can estimate correctly the original labels, it is remarkable because is normal that evaluators, i.e. doctors, can label correctly a one class, the normal or control class, the pathologic class contains the labeling errors. For the last reason, the present methodology is valid and present an reliable alternative to address multilabeler problems.

For future works, we are aiming to explore alternatives to improve the reference labeling vector setting, since the majority vote may not be an adequate reference for all cases, specially, when there are many supposed wrong labelers. Aditionally, it is necessary review other ways to penalize the error labels oriented to improve the penalty factor in terms of the natural distribution of the data. Finally, the extension of this methodology to the multiclase case is other outstanding contribution to the state of the art.

Acknowledgments. 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 and Universidad Nacional de Colombia - Manizales and the project "Servicio de Monitoreo Remoto de Actividad Cardíaca para el Tamizaje Clínico en la red de Telemedicina del Departamento de Caldas" financed with the "Fondo Estampilla Universidad Nacional de Colombia Manizales - Universidad de Caldas".

References

1. Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. Journal of Machine Learning Research 11, 1297–1322 (2010)
2. Dekel, O., Shamir, O.: Vox populi: Collecting high-quality labels from a crowd. In: Proceedings of the 22nd Annual Conference on Learning Theory (2009)
3. Smyth, P., Fayyad, U.M., Burl, M.C., Perona, P., Baldi, P.: Inferring ground truth from subjective labelling of venus images. In: NIPS 1994, pp. 1085–1092 (1994)
4. Crammer, K., Kearns, M., Wortman, J., Bartlett, P.: Learning from multiple sources. In: Advances in Neural Information Processing Systems, vol. 19 (2007)
5. Dekel, O., Shamir, O.: Good learners for evil teachers. In: ICML, p. 30 (2009)
6. Frank, A., Asuncion, A.: UCI machine learning repository (2010)