ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

A multi-class extension for multi-labeler support vector machines

D.H. Peluffo-Ordóñez[1] S. Murillo-Rendón[2] J. D. Arias-Londoño[3] G. Castellanos-Domínguez[2] *

1- Université catholique de Louvain, Machine Learning Group - ICTEAM
2- Universidad Autónoma de Manizales, Grupo de Ingeniería de Software
3- Universidad de Antioquia, Department of Systems Engineering
4- Universidad Nacional de Colombia, Signal Processing and Recognition Group

Abstract. In recent years, there has been an increasing interest in the design of pattern recognition systems able to deal with labels coming from multiple sources. To avoid bias during the learning process, in some applications it is strongly recommended to learn from a set of panelists or experts instead of only one. In particular, two aspects are of interest, namely: discriminating between confident and unconfident labelers, and determining the suitable ground truth. This work presents an extension of a previous work, which consists of a generalization of the two class case via a modified one-against-all approach. This approach uses modified classifiers able to learn from multi-labeler settings. This is done within a soft-margin support vector machine framework. Proposed method provides ranking values for panelist as well as an estimate of the ground truth.

1 Introduction

In supervised machine learning approaches, typically, the data labels are known in advance. Nonetheless, in some cases the labeling process is not provided by a unique expert labeler (panelist), therefore, there is no a clearly, identified ground truth. To deal with this matter, a multi-labeler strategy should be performed, allowing for reducing the influence of wrong labels regarding the assumed ground truth. Since there is uncertainty on the accuracy or confidence of the panelist team, the multi-labeler analysis is aimed to compensate the negative effect of wrongly labeled samples. Such compensation may either improve the learning process, in terms of classification [1], or yield penalty factors for quantifying the panelist's efficiency [2]. The latter case, is commonly used for discriminating between confident and unconfident labelers, specially, when ground truth is a determinant starting point and/or querying experts may be significantly expensive [3]. Particularly, support vector machines (SVM) have shown to be a suitable alternative to deal with this issue, mainly due to its versatility [1, 4–6].

Previous works have addressed the multi-labeler learning problem by modifying the formulation of binary-SVM-based classifiers [4, 7]. Such classifiers incorporate a the panelist's labeling within a soft-margin SVM approach (SM-SVM). In particular, a least-squares formulation is posed leading to a quadratic problem which can be solved by conventional quadratic programming algorithms. In this this work, we present a natural extension of such a binary case to a multi-class version by means of a modified

*This is work is funded by the project "Universidad Nacional de Colombia, Grupo de control y procesamiento digital de señales" código 20501007205.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

one-against-all strategy, which keeps the effect of multi-labeler within the decision process. Doing so, our approach is able to provide penalty factors for panelists regarding the accuracy on classifying each single class (one penalty factor per class). To asses the performance of our method, a controlled artificial data set is used which allows for analyzing how our method behaves on different levels of separability as well as different noise levels on labelers. Furthermore, obtained penalty factors are compared with standard supervised measures. Results show that our method is able to asses the concordance among panelists involving the structure of data.

This paper is organized as follows: Section 2 outlines briefly the binary approach for multilabeler analysis. Section 3 describes our proposed multi-class extension. Section 4 holds some results and discussion. Finally, final remarks are presented in 5.

2 Bi-class multi-labeler classifier based on SVM (BMLC)

The binary approach for multi-labeler settings studied here consists of an extension of a two-class based on SVM classifier to multi-labeler analysis. Such an approach, here termed BMLC, was recently introduced in [7]. BMLC works as follows: For the given data matrix $\boldsymbol{X} \in \mathbb{R}^{m\times d}$ and labeling vector $\boldsymbol{y} \in \mathbb{R}^m$ (being d the number of considered features and m the number of samples), the ordered pair $\{\boldsymbol{x}_i, y_i\}$ is defined 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, such that $y_i \in \{1, -1\}$. Furthermore, a hyperplane model is assumed in the form: $e_i = \boldsymbol{w} \cdot \boldsymbol{x}_i + b = \boldsymbol{w}^\top \boldsymbol{x}_i + b$, where $\boldsymbol{w}$ is an orthogonal vector and b is a bias term. For classification purposes, function $f(\boldsymbol{x}) = \text{sign}(e_i)$ can be used as a decision criterion. Then, we can write soft-margin SVM formulation incorporating a slack variable ξ_i, such that: $1 - y_i e_i;\ \forall i$, where $b = 0.$, as follows;

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

where $\boldsymbol{\xi} \in \mathbb{R}^m = [\xi_1, \ldots, \xi_m]$ and $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. Since in (1) term $1 - y_i e_i$ becomes upper bounded by ξ_i, one can infer that minimizing $f(\cdot)$ with regard to $\boldsymbol{w}$ should be the same as minimizing $\hat{f}(\cdot)$ regarding both $\boldsymbol{w}$ and ξ_i.

To address the multi-labeler problem, we aim to design a suitable supervised classifier involving the information given by different sources (that is, labeling vectors). To this end, a penalty factor θ_t is usually incorporated, such that $\hat{f}(\cdot)$ decreases when adding correct labels. Otherwise it should increase. In [1], the SM-SVM formulation of Eq. (1) is also suggested, but the formulation is relaxed to its linear version. Instead, this work proposes to extend SV-SMV in the genuine quadratic version to improve the estimation of $\boldsymbol{w}$. Particularly, given a set of k panelists assigning their corresponding labeling vectors, each t-th panelist is to be associated to the penalty factor θ_t, where $t \in [k]$ and $[k] = \{1, \ldots, k\}$. Accordingly, by including the penalty factor vector $\boldsymbol{\theta} = [\theta_t]_{1\leq t\leq k}$, we can re-write the functional given in Eq. (1), as follows:

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

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

where c_{it} is the coefficient (weight) for the linear combination of all θ_t, which represents the relevance of the information given by t-th panelist over the sample i. Thus, the term $(1/k)\sum_{t=1}^{k} c_{it}\theta_t$ is a weighted average of the whole set of penalty factors. Defining an auxiliary variable $\hat{\xi}_i$ in the form:

$$\hat{\xi}_i = \xi_i + \frac{1}{k}\sum_{t=1}^{k} c_{it}\theta_t \Rightarrow \hat{\boldsymbol{\xi}} = \boldsymbol{\xi} + \frac{1}{k}\boldsymbol{C}\boldsymbol{\theta}, \tag{3}$$

the problem (2) becomes $\min(\lambda/2)\boldsymbol{w}^\top\boldsymbol{w} + (1/2m)\hat{\boldsymbol{\xi}}^\top\hat{\boldsymbol{\xi}}$; s.t. $\hat{\boldsymbol{\xi}} \geq \mathbf{1}_m - (\boldsymbol{X}\boldsymbol{w}) \circ \boldsymbol{y}$. Then, to solve the problem, first the corresponding Lagrangian is formed. Second, we pose a dual problen by solving the Karush-Kuhn-Tucker (KKT) conditions for the Lagrangian, and eliminating primal variables by expressing them depending only on the dual variable, constants, and known variables. This procedure is detailed in a previous work [7]. Such a dual formulation is a quadratic problem with linear constraints easy to be solved by a quadratic programming procedure. Following from this, vector $\boldsymbol{w}$ is found, and then recalling equation 3 penalty factors can be directly calculated by

$$\boldsymbol{\theta} = \boldsymbol{C}^\dagger(\mathbf{1}_m - (\boldsymbol{y} \circ \boldsymbol{e}) - \boldsymbol{\xi}), \tag{4}$$

where $\boldsymbol{C}^\dagger$ is the pseudo-inverse matrix of $\boldsymbol{C}$. As an important characteristic of our method, it is worth to mention that despite being, in principle, designed to calculate the influence of each labeller, BMLC also outputs a suitable classifier modeled by $\boldsymbol{e} = \boldsymbol{X}\boldsymbol{w} + b\mathbf{1}_m$. This is approach is detailed in [4].

3 Multi-class extension

This section is aimed to deal with various classes ℓ, such that $\ell \in \{1, \ldots, c\}$, where c is the amount of considered classes. When dealing with problems involving more than two classes, a multi-class extension to the formulation described in previous section is needed. The first intuition to carry out this task is perhaps by using traditional multi-class approaches. Among this kind approaches, we find one against one, directed acyclic graph and one against all (OaA), which in general reach similar performance [8, 9]. In particular, within a OaA-based framework a multi-class SM-SVM formulation for multi-labeler analysis is introduced. This approach consists of building a number of SVM models, one per class. Applying c times the bi-class approach described in section 2, a multi-class approach is accomplished. In general, in case of using SVM-approaches, class c is compared with the remaining ones in such a way that it is matched with a positive label, meanwhile the others with a negative label [8]; so that a binary labeling vector per each single class is formed. Concretely, the labeling reference vector $\bar{\boldsymbol{y}}^{(\ell)}$ associated to class ℓ is assumed in the form: $\bar{y}_i^{(\ell)} = 1$ if $\boldsymbol{x}_i$ belongs to class ℓ, otherwise 0. In this sense, the proposed approach stated in (1) can be generalized as:

$$\min_{\boldsymbol{w}^{(\ell)},\boldsymbol{\xi}^{(\ell)}} \frac{\lambda_\ell}{2}\|\boldsymbol{w}^{(\ell)}\|^2 + \frac{1}{2m}\sum_{i=1}^{m}\left(\xi_i^{(\ell)} + \frac{1}{k}\sum_{t=1}^{k} C_{it}^{(\ell)}\theta_t^{(\ell)}\right)^2 \tag{5}$$

$$\text{s.t.} \quad \xi_i^{(\ell)} \geq 1 - \bar{y}_i^{(\ell)} e_i^{(\ell)} - \frac{1}{k}\sum_{t=1}^{k} C_{it}^{(\ell)}\theta_t^{(\ell)}.$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Consequently, the decision hyperplanes are given by $\{e_i^{(\ell)}, \ldots, e_i^{(\ell)}\}$, where $\boldsymbol{e}^{(\ell)} = \boldsymbol{X}\boldsymbol{w}^{(\ell)} + b^{(\ell)}\mathbf{1}_m$. Once vectors $\boldsymbol{w}^{(\ell)}$ are calculated by solving BMLC, corresponding penalty factors $\boldsymbol{\theta}^{(\ell)}$ are calculated with (4). Finally, global penalty factors $\bar{\boldsymbol{\theta}} \in \mathbb{R}^k$ can be estimated by averaging those obtained per each class

$$\bar{\boldsymbol{\theta}} = \frac{1}{c}\sum_{\ell=1}^{c} \boldsymbol{\theta}^{(\ell)}. \tag{6}$$

As BMLC, the set of vectors $\boldsymbol{w}^{(\ell)}$ from the proposed multi-class extension also provides a reliable labeling vector $\hat{\boldsymbol{y}} \in \{1, \ldots, c\}^m$ given by: $\hat{y}_i = \arg\max_{\ell \in \{1,\ldots,c\}} e_i^{(\ell)}$.

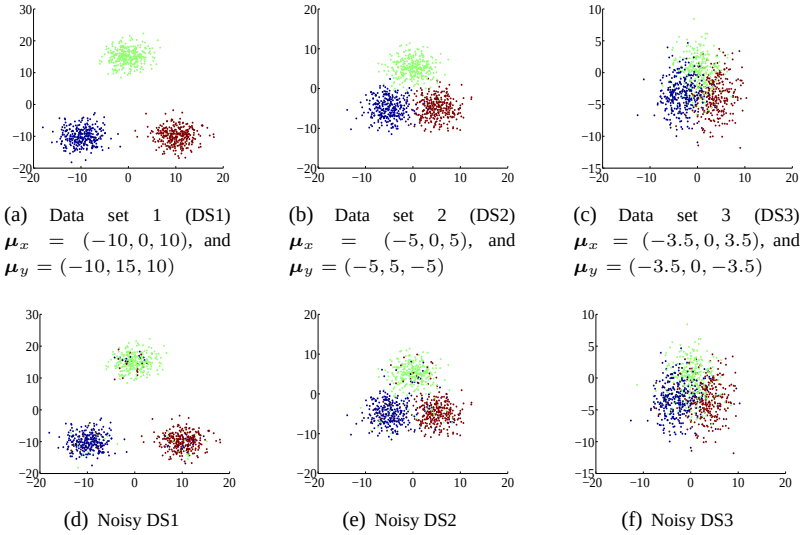
4 Results and discussion

For experimental results, we consider three artificial data sets corresponding to two-dimensional Gaussian clouds. We simulate 3 classes which are ruled by the mean vectors whose XY position are $\boldsymbol{\mu}_x = [\mu_x(\ell)]$ and $\boldsymbol{\mu}_y = [\mu_y(\ell)]$, as well as the standard deviation vector $\boldsymbol{\sigma} = [\sigma(\ell)]$, where $\ell \in \{1, \ldots, c\}$. By design, all the classes are built identically distributed by setting the standard deviation as $\sigma(\ell) = 2.5, \forall \ell$. Data are simulated in such a manner that the assigned labeling or ground truth is known and the class separability can be modified. Also, in order to test the sensibility to wrong labels of our method, we simulate noisy labels by adding a percentage ϵ of error. In this case, noisy labels are those that have been replaced by any of the other classes altering then the accuracy of labeling. Here, we introduce the same error ϵ to the three classes by randomly replacing labels with wrong ones. In the top row of Fig. 1, scatter plots of simulated data sets regarding the assigned label (ground truth) are depicted. Likewise, in the below row, examples of noisy data with $\epsilon = 10\%$ are shown. To quantify the accuracy of the resultant labels regarding the ground truth, two performance measures are used: the normalized mutual information (NMI) [10] and the adjusted random index (ARI) [11]. These measures are also applied to compare our method with simple multi-labeler approaches consisting of obtaining a single labeling vector by calculating either the mode (majority vote) or the average of the labels for each sample. In addition, in order to evaluate how stable is our approach, the procedure is iterated 20 times, and the values of mean and standard deviation are calculated. Overall results are presented in Table 1.

In general, we can notice that our method outperforms traditional procedures. Despite of the presence of wrong labels, decision functions are well performed since the structure of data is taken into account within the hyperplane model. Even when the noisy labeling error is increased, proposed approach output remains a suitable estimate of the ground truth as can be seen in Table 1. Nonetheless, when classes are very close to each other or having overlapping as happens in SD 3, our method does not reach that high performance. This fact can be attributed to the model which is linear. Indeed, it can be seen as kernel SVM with a linear kernel. Although our method is multi-class, it is limited to linear function decisions. To overcome this drawback, the number of classes can be slightly increased to classify overlapped samples as new classes.

In Fig. 2, the resulting penalty factors at different errors ϵ (0, 10, 20, 30) are depicted. Since the labelers are created with the same error ϵ, one can expect the same

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.



(a) Data set 1 (DS1) $\boldsymbol{\mu}_x = (-10, 0, 10)$, and $\boldsymbol{\mu}_y = (-10, 15, 10)$

(b) Data set 2 (DS2) $\boldsymbol{\mu}_x = (-5, 0, 5)$, and $\boldsymbol{\mu}_y = (-5, 5, -5)$

(c) Data set 3 (DS3) $\boldsymbol{\mu}_x = (-3.5, 0, 3.5)$, and $\boldsymbol{\mu}_y = (-3.5, 0, -3.5)$

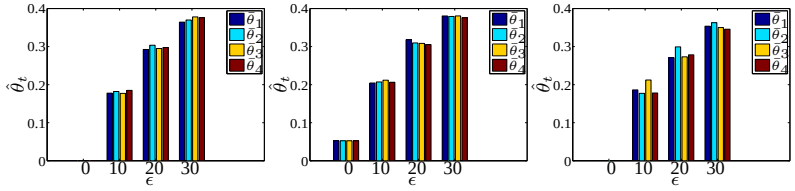
(d) Noisy DS1

(e) Noisy DS2

(f) Noisy DS3

Fig. 1: Examples of two dimensional three-class Gaussian data. The mean of each class is given by $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_x$, meanwhile its ratio given by a value of standard deviation ($\sigma(\ell) = 2.5$)

penalty factors for each labeler. Indeed, the values of $\bar{\theta}_t$ are roughly constant for each considered error ϵ. Nonetheless, it is noteworthy that in case of less separability (DS 3), penalty factors fluctuate more. This fact can be attributed to the presence of samples in the ambiguity region, i.e. where classifiers assign classes at random or where there is a mismatch between decision functions in the multi-labeler extension process.



(a) DS 1

(b) DS 2

(c) DS 3

Fig. 2: Averaged penalty factors $\bar{\theta}_t$ for considered data sets at different errors ϵ.

5 Conclusions and future work

Generally, the use of multi-labeler strategy may provide a better design and training of classifiers in comparison with one-labeler approaches. Experimentally, we prove that the proposed approach is capable to quantify the confidence of a set of panelists taking into consideration the natural structure of data. The here proposed extension to deal with multi-class settings shows to keep the same benefits of binary case such as

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Data set	ϵ (%)	Proposed method		Mode		Average	
		NMI	ARI	NMI	ARI	NMI	ARI
DS1	0	1±0.00	1±0.00	1	1	1	1
	10	**0.95±0.01**	**0.97±0.00**	0.92	0.96	0.78	0.77
	20	**0.78±0.01**	**0.82±0.00**	0.66	0.74	0.52	0.48
	30	**0.62±0.00**	**0.67±0.01**	0.38	0.44	0.37	0.29
DS2	0	0.9±0.00	0.93±0.00	1	1	1	1
	10	**0.84±0.01**	**0.89±0.01**	0.79	0.83	0.73	0.73
	20	**0.68±0.00**	**0.75±0.01**	0.62	0.70	0.51	0.44
	30	**0.52±0.01**	**0.57±0.00**	0.36	0.42	0.35	0.28
DS3	0	0.99±0.00	0.99±0.00	1	1	1	1
	10	**0.96±0.01**	**0.93±0.00**	0.90	0.93	0.73	0.73
	20	0.62±0.00	0.67±0.00	**0.66**	**0.72**	0.54	0.48
	30	0.4±0.01	0.43±0.00	**0.42**	**0.48**	0.34	0.26

Table 1: NMI and ARI for considered data sets. Our method is iterated 10 times to assess its stability. Proposed method is compared with mode and average procedures at different introduced errors ϵ.

the flexible learning process, and the capability to penalize labellers. In addition, our approach is able to deal with moderate noisy labels keeping a good performance.

For future work, we are aiming at exploring alternatives to improve the reference labeling vector setting, since the simple majority vote may not be adequate as a reference, specially, when there are many supposed wrong labelers. Also, kernelized version can be devoloped.

References

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

[2] Pinar Donmez, Jaime G Carbonell, and Jeff Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 259–268. ACM, 2009.

[3] Wei Wang and ZhiHua Zhou. Learnability of multi-instance multi-label learning. *Chinese Science Bulletin*, 57(19):2488–2491, 2012.

[4] S Murillo, D. H. Peluffo, and G Castellanos. Support vector machine-based approach for multi-labelers problems. *European Symposium on Artificial Neural Networks, Computational Inteligence and Machine Learning*, 2013.

[5] Yu Zhang and Dit-Yan Yeung. Multilabel relationship learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(2):7, 2013.

[6] Ricardo Cerri, André Carlos PLF de Carvalho, and Alex A Freitas. Adapting non-hierarchical multilabel classification methods for hierarchical multilabel classification. *Intelligent Data Analysis*, 15(6):861–887, 2011.

[7] S Murillo-Rendón, D Peluffo-Ordóñez, Julián D Arias-Londoño, and CG Castellanos-Domínguez. Multi-labeler analysis for bi-class problems based on soft-margin support vector machines. In *Natural and Artificial Models in Computation and Biology*, pages 274–282. Springer, 2013.

[8] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[9] Yi Liu and Yuan F Zheng. One-against-all multi-class svm classification using reliability measures. In *IEEE International Joint Conference on Neural Networks*, volume 2, pages 849–854. IEEE, 2005.

[10] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[11] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 1(2):193–218, 1985.