Gaussian Processes Regression with Multiple Annotators: When the Annotator Performance Is Not Homogeneous

Julián Gil González(✉), Andrés Marino Álvarez, and Álvaro Angel Orozco

Faculty of Engineering, Universidad Tecnológica de Pereira, 660003 Pereira, Colombia
{jugil,andres.alvarez1,aaog}@utp.edu.co

Abstract. In supervised learning problems, the right label (also known as the gold standard or the ground truth) is not available because the label acquisition can be expensive or infeasible. Instead of that gold standard, we have access to some annotations provided by multiple annotators with different levels of expertise. Hence, trivial methods such as majority voting (or average in regression problems) are not suitable since they assume homogeneity between the expertise of the labelers. In this work, we introduce a regression approach based on Gaussian processes, where we consider that the expertise of the labelers is non-homogeneous across the input space–(GPR-MANH). The idea is to assume that the input space can be represented by a defined number of regions where each annotator exhibit a particular level of expertise. Experimental results show that our methodology can estimate the performance of annotators even if the gold standard is not available, defeating state-of-the-art techniques.

1 Introduction

A typical supervised learning scenario comprises the computation of a function, which maps from inputs (samples) to outputs (labels), where it is assumed that exists an oracle who gives the correct label (also known as ground truth or gold standard) for each sample in the training set [1]. However, in many real-world applications, the gold standard is not available, because the process to acquire it is expensive, unfeasible, time-consuming or the label corresponds to a subjective assessment [2]. Instead of the ground truth, it is possible to access several labels provided by multiple annotators or sources. This information can be acquired using web sources, crowdsourcing platforms or the opinion of multiple experts. For instance, social networks (e.g., Twitter, Facebook) can be used to obtain information about a specific problem such as product rating or sentiment analysis [3]. Likewise, in problems where the gold standard is not available, we can use platforms like Amazon Mechanical Turk (AMT), LabelMe, Crowdflower.[1]

[1] www.mturk.com; labelme2.csail.mit.edu/; crowdflower.com.

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 91–99, 2019.
https://doi.org/10.1007/978-3-030-13469-3_11

This kind of platforms offers a cost-effective, and efficient way to obtain labeled data [4]. On the other hand, in problems of computer-aided diagnosis, we can obtain subjective assessment provided by different experts [5,6]. Nevertheless, the information collected from these multiple sources could be subjective, noisy or even misleading [6]. Trivial solutions to deal with multiple labelers scenarios include *(i) to consider as the gold standard the output from one of the labelers*, and *(ii) to assume the majority voting (or the average in the case of regression) from the annotations as an estimation for the ground truth*. However, these approaches are not suitable due to they assume homogeneity between the performance of the annotators [2].

On the other hand, *Learning from crowds* is a particular area of supervised learning, which deals with different machine learning paradigms in the presence of multiple annotators, including classification, sequence labeling, and regression. Among the methodologies developed in the area of learning from crowds, we can identify two main groups. The first group named label aggregation are focused only on estimating the gold standard, which is then used to train a supervised learning scheme. On the other hand, the second group comprises the works that are focused on training supervised learning models directly from the labels of multiple sources. Regarding the classification paradigm, we recognize the approach proposed in [6], which comprises the estimation of the annotator expertise (in terms of sensitivity and specificity) through a maximum likelihood-based approach from repeated responses (labels). In this sense, this model estimate jointly the gold standard and the classifier parameters using a logistic regression-based framework. Similarly, the authors in [7] propose an extension of the work proposed in [6] aiming to introduce a Gaussian processes model as the classification scheme. On the other hand, with respect to real-valued label (i.e. Regression models), the authors in [1] propose a Gaussian processes model to deal with multiple annotators, where the performance of the labelers is coded by including a per-annotator variance in the likelihood function–(GPR-MAH). However, they assume that the labeler performance is homogeneous across the input space, which is a weak assumption as was demonstrated in [8]. The above assumption was relaxed by the work in [9]. This approach codes the performance using a Gaussian process model, which estimates the annotators expertise as a non-linear function of the gold standard and the input space.

In this work, we present a regression approach based on Gaussian processes, where the expertise of the labelers is non-homogeneous across the input space–(GPR-MANH). Our approach follows the idea of GPR-MAH, in the sense that we use a Gaussian processes method to model the regression function and assign a per-annotator variance to capture the performance of the labelers. However, unlike GPR-MAH, our methodology relaxes the assumption that the performance of each annotator is homogeneous across the input space by considering that the input space can be represented by a number specific of clusters, where each annotator exhibits different performances. We empirically show, using simulated annotators, that our methodology can be used to learn regression models using noisy data from multiple sources, outperforming state-of-the-art techniques. The remainder of this

paper is organized as follows. Section 2 describes the background of our approach. Sections 3 and 4 present the experiments and discuss the results obtained. Finally, Sect. 5 outlines the conclusions and future work.

2 Probabilistic Formulation

A regression scenario has the primary goal to estimate a function $f : \mathscr{X} \rightarrow \mathscr{Z}$ using a training set $\{\mathbf{x}_n, z_n\}_{n=1}^{N}$, where $\mathbf{x}_n \in \mathscr{X} \subseteq \mathbb{R}^P$ is a $P-$dimensional input feature vector corresponding to the $n-$th instance with output $z_n \in \mathscr{Z} \subseteq \mathbb{R}$. In a typical regression configuration, each sample $\mathbf{x}_n$ is assigned to a single output z_n, i.e., the ground truth. However, in many real-world regression problems instead of the ground truth we have multiple labels provided by R sources with different levels of expertise [1]. Moreover, we assume that each annotator annotates $N_r \leq N$ observations. In this sense, it is possible to build a data set for the annotator $r \in \{1, 2, \ldots, R\}$, $\mathscr{D}_r = \{\mathbf{X}_r, \mathbf{y}_r\}$, where $\mathbf{X}_r \in \mathbb{R}^{N_r \times P}$ and $\mathbf{y}_r \in \mathscr{Y}_r \subseteq \mathbb{R}$ are the input feature matrix and the labels given by the r-th annotator, respectively. Besides, $\mathbf{X}_r$ holds row vectors $\mathbf{x}_n^r$ and $\mathbf{y}_r$ is composed by elements y_n^r, where y_n^r is the m-th annotation of sample $\mathbf{x}_n^r$. Now given the data set from multiple annotators $\mathscr{D} = \left\{\mathbf{X} = \cup_{r=1}^{R} \mathbf{X}_r, \mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_R\}\right\}$, our goals are: First, to estimate the unknown gold standard for the instances in the training set $\mathbf{z} = [z_1, \ldots, z_N]$. Second, to compute the performance of the labelers as a function of the ground truth and the input space. Finally, the third objective is to build a regression model based on Gaussian processes which generalizes well on unseen data.

Concerning this, we follow the model for the labels proposed in [1], $y_n^r = z_n + \mathscr{N}(0, \sigma_r^2)$, where they consider that the parameter σ_r^2 (related to the performance of the r-th annotator) is homogeneous across the input space. However, as we established previously, the principal aim of our work is to model the annotator expertise based on the assumption that it is no-homogeneous across the input space. For doing so, we assume that the input space $\mathscr{X}$ can be represented using K clusters based on the input space Euclidean distances, where each annotator exhibits a particular performance. Accordingly, the model proposed for the labels y_n^r follows $y_n^r = z_n + \mathscr{N}(0, (\sigma_k^r)^2)$, where $(\sigma_k^r)^2 \in \mathbb{R}^+$ is the variance for the r-th labeler in the cluster $k \in \{1, 2, \ldots, K\}$. Assuming independence between annotators, and the fact that each annotator labels $\mathbf{x}_n$ independently, the likelihood is given as follows

$$p\left(\mathbf{Y}|\mathbf{z}\right) = \prod_{k} \prod_{n \sim k} \prod_{r \sim n} \mathscr{N}\left(y_n^r | z_n, \left(\sigma_k^r\right)^2\right) = c\mathscr{N}\left(\hat{\mathbf{y}}|\mathbf{z}, \hat{\boldsymbol{\Sigma}}\right), \tag{1}$$

where $c \in \mathbb{R}$ is independent of $\mathbf{z}$, the diagonal matrix $\hat{\boldsymbol{\Sigma}} \in \mathbb{R}^{N \times N}$ has elements $\hat{\sigma}_{nk}^2$, the vector $\hat{\mathbf{y}} \in \mathbb{R}^N$ has entries $\hat{y}_{nk}$. Also, $\hat{\sigma}_{nk}^{-2} = (\sum_{r \sim n} 1/(\sigma_k^r)^2)^{-1}$, $\hat{y}_{nk} = \hat{\sigma}_{nk}^2 \sum_{r \sim n} y_n^r/(\sigma_k^r)^2$. The notation $r \sim n$ refers to "take into account only the labelers who annotated the n-th observation" and $n \sim k$ indicates the sample n belonging to the k-th cluster. Assuming a Gaussian process prior for $\mathbf{z}$ given as $p(\mathbf{z}) = \mathscr{N}(\mathbf{z}|\mathbf{0}, \mathbf{K})$, with kernel matrix $\mathbf{K}$ computed using a particular kernel

function $k : \mathbb{R}^P \times \mathbb{R}^P \to \mathbb{R}$, the posterior over the latent variable $\mathbf{z}$ is computed as follows $p(\mathbf{z}|\mathbf{Y}, \mathbf{X}) = \mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})$, where $\mathbf{m} = (\mathbf{K}^{-1} + \hat{\boldsymbol{\Sigma}}^{-1})^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{y}}$, and $\mathbf{V} = (\mathbf{K}^{-1} + \hat{\boldsymbol{\Sigma}}^{-1})^{-1}$. In turn, it can be shown that the posterior over a new observation $f(\mathbf{x}_*)$ follows

$$p(f(\mathbf{x}_*)|\mathbf{Y}) = \mathcal{N}(f(\mathbf{x}_*)|\bar{f}(\mathbf{x}_*), k(\mathbf{x}_*, \mathbf{x}'_*)), \tag{2}$$

where $\bar{f}(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{X})(\mathbf{K} + \hat{\boldsymbol{\Sigma}})^{-1}\hat{\mathbf{y}}$ and $k(\mathbf{x}_*, \mathbf{x}'_*) = k(\mathbf{x}_*, \mathbf{x}'_*) - k(\mathbf{x}_*, \mathbf{X})(\mathbf{K} + \hat{\boldsymbol{\Sigma}})^{-1}k(\mathbf{X}, \mathbf{x}'_*)$. The free parameters related to the model (the hyper-parameters of the kernel function, and the variances associated to the annotators in each region) are estimated by optimizing the negative log of the evidence, which is given as

$$\begin{aligned} -\log p(\mathbf{Y}) = \frac{1}{2}\log|\mathbf{K} + \hat{\boldsymbol{\Sigma}}| + \frac{1}{2}\hat{\mathbf{y}}^\top(\mathbf{K} + \hat{\boldsymbol{\Sigma}})^{-1}\hat{\mathbf{y}} - \frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}| \\ + \frac{1}{2}\sum_k \sum_{n\sim k} \sum_{r\sim n} \frac{(y_n^r)^2}{(\sigma_k^r)^2} - \frac{1}{2}\sum_k \sum_{n\sim k} \frac{\hat{y}_{nk}^2}{\hat{\sigma}_{nk}^2} - \sum_k \sum_{n\sim k} \sum_{r\sim n} \log \frac{1}{\sigma_k^r} + \frac{\zeta}{2}\log 2\pi, \end{aligned}$$

where $\zeta = \sum_{r=1}^{R} N_r$. To summarize, we propose a regression scheme with multiple annotators based on Gaussian processes, where the performance of the annotators is coded by including a per-annotator variance in the likelihood function. Unlike GPR-MAH, we assume that the input space is represented by K regions, where the annotators exhibit a particular performance, which is represented by a variance $(\sigma_k^r)^2$.

3 Experimental Set-Up

Testing Datasets. To test our GPR-MANH, we use three datasets for regression of the well-known *UCI repository*[2]. The used datasets include: *Auto MPG*–(Auto), *Concrete Compressive Strength*–(Concrete), and *Boston Housing Data*–(Housing). The above datasets were chosen based on state-of-the-art works [1,9].

Simulated Annotations. The datasets from the UCI repository are mainly focused on supervised learning without multiple sources. Thus, we establish two methods for simulating multiple annotators: *(i)* Homogeneous Gaussian noise [1], that samples a random number $\varepsilon_n^r \in \mathbb{R}$ from a Gaussian distribution with zero mean and variance $\tau_r^2 \in \mathbb{R}^+$; then the annotations are simulated as, $y_n^r = z_n + \varepsilon_n^r$. Accordingly, τ_r^2 codes the performance of the annotators, the higher is its value, the lower the expertise level of the r-th labeler. *(ii)* Non-homogeneous Gaussian noise [8]. This simulation approach comprises the following steps: First, we split the data into L clusters using the k-means algorithm. Next, the annotations

[2] http://archive.ics.uci.edu/ml.

given by the r-th annotator for samples in the l-th cluster follows, $y_{nl}^r = z_n + \mathcal{N}\left(0, \lambda_{lr}^2\right)$, where $\lambda_{lr}^2 \in \mathbb{R}^+$ codes the labeler expertise in the region l. Hence, we simulate labelers where its expertise varies depending on the input space.

Validation Approaches and Learning Assessments. Aiming to validate the performance of our approach, we take into account the following state-of-the-art models. *(i) Gaussian Process-based regression with majority voting*–(GPR-Av), where a typical regression model is trained using as the gold standard the average from the annotations. The kernel hyperparameters related to this Gaussian processes are estimated by optimizing the marginal likelihood [11]. *(ii) Learning from Multiple Observers with Unknown Expertise*–(LMO), that uses a Gaussian process to code the expertise of the labelers as a function of the gold standard and the input samples. The parameter estimation is carried out using a Maximum a Posterior (MAP) approach [9]. *(iii) Learning from Multiple Annotators with Gaussian Processes*–(GPR-MA), where a per-annotator variance is included in the likelihood function to capture the information from multiple annotators. The hyperparameters related to the kernel function and the variances of each annotator are estimated by minimizing the minus log of the evidence.

Furthermore, the validation is carried out by estimating the regression performance in terms of the mean squared error (note that we have access to the gold standard). A cross-validation scheme is carried out with 30 repetitions (70% of the samples as training and 30% as testing).

4 Results and Discussions

First, we perform a controlled experiment aiming to verify the capability of our approach for dealing with regression setting in the context of multiple sources. For this first experiment, the training samples $\mathbf{X}$ are generated by randomly selecting 60 points in the interval $[0, 1]$, and the ground truth is computed as $z_n = \sin(2\pi x_n)\sin(6\pi x_n)$. The instances for testing are formed with 600 equally spaced samples from the interval $[0, 1]$. We simulate three labelers with different levels of expertise by using the simulation methods described in Sect. 3. For the "Homogeneous Gaussian noise" we use $\boldsymbol{\tau} = (0.25,\ 0.5,\ 0.75)$. On the other hand, for the"Non-homogeneous Gaussian noise", we split the input space into three regions and use the following parameters:

$$\boldsymbol{\Lambda} = \begin{pmatrix} 0 & 0.65 & 1.0 \\ 0.25 & 0 & 0.75 \\ 0.1 & 0.75 & 0 \end{pmatrix}.$$

Here, the matrix $\boldsymbol{\Lambda}$ is formed by elements λ_{lr}^2, which indicates the variance for the r-th annotator in the cluster l. For testing our approach, we use a clustering algorithm based on *affinity propagation* [12] aiming to obtain a proper representation of the input space $\mathscr{X}$. Similarly, for the Gaussian processes model, the kernel is fixed as a squared exponential function [11]. Figure 1 shows a visual comparison among the performance of our GRP-MANH and the methods considered for validation (GPR-Av, LMO, GPR-MA), considering the case when the

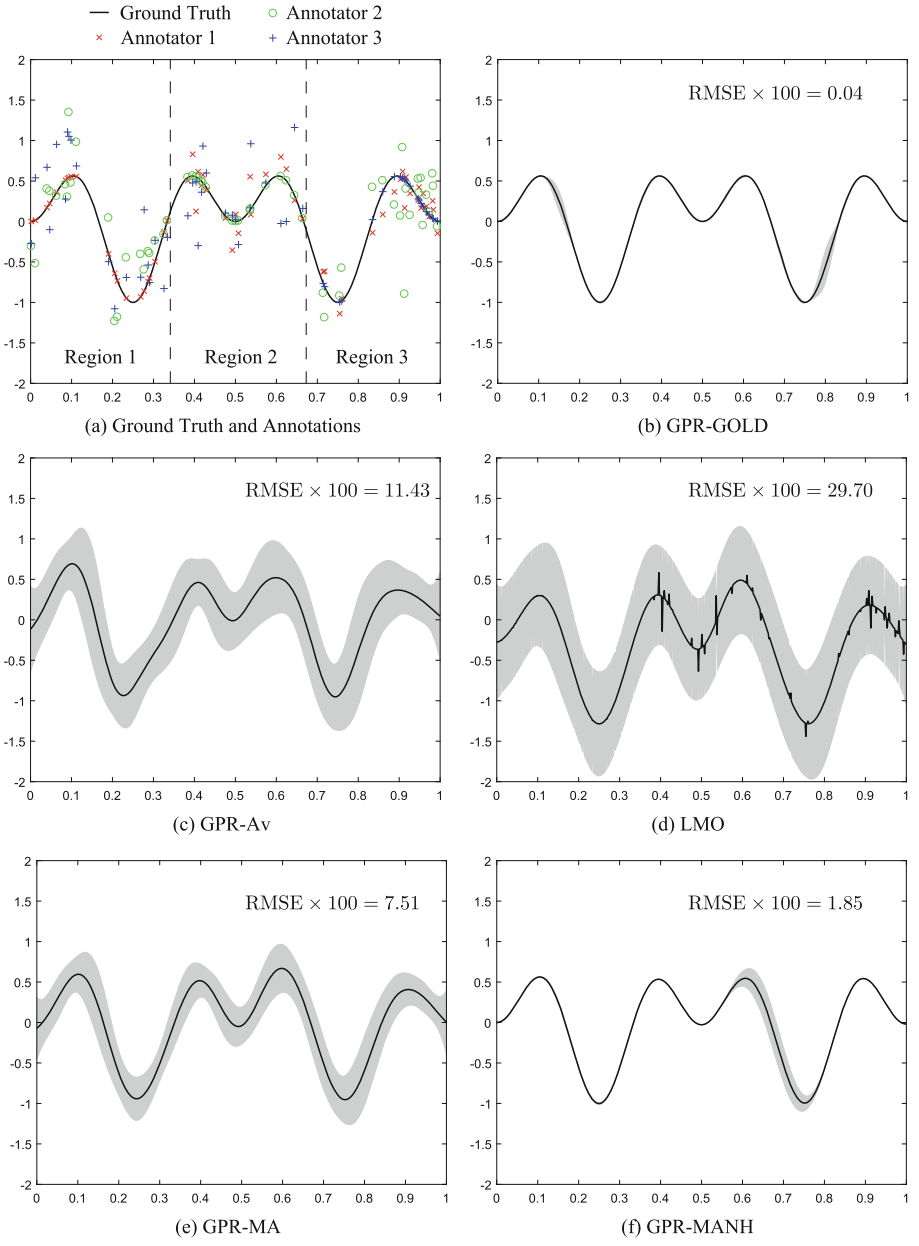

Fig. 1. Results for the first experiment. In (a) we expose the ground truth and the synthetic annotations, which are generated using the simulation method "Non-homogeneous Gaussian noise". In (b), (c), (d), (e), and (f) we respectively show the regression results for GPR-GOLD, GPR-Av, LMO, GPR-MA, and GPR-MANH. Shaded areas represent the variance for the predictions.

data from multiple annotators are generated using "Non-homogeneous Gaussian noise". Remarkably, we note that our approach can perform regression settings in scenarios where the gold standard is not available, and the expertise of the annotators is not homogeneous across the input space. In fact, it is possible to observe that the uncertainty of the predictions of our approach is remarkably lower when compared with the validation methodologies. The above can be explained in the sense that our GRP-MANH can perform a better codification of the annotator expertise.

Now, we carry out regression experiments using three datasets from the UCI repository, where we simulate three annotators with different levels of expertise using the simulation parameters described below. Table 1 reports the mean and the standard deviation for the root mean squared error–(RMSE) predicted. Besides, the method with the highest performance is highlighted in bold, excluding the upper bound (GPR-GOLD), which is a Gaussian Processes for regression trained with the true labels. As seen, most of the regression methods from multiple annotators considered in this work (GPR-MA, and GPR-MANH) outperform the average baseline (GPR-av) in most cases, which is not surprising, since this baseline does not consider differences between the expertise of the labelers. Furthermore, we empirically demonstrated that our approach is not affected where the performance of the annotators is not homogeneous across the input space. In fact, our GRP-MANH outperforms all the models considered in this work for validation under the two methods used for generating the synthetic annotations (homogeneous Gaussian noise and non-homogeneous Gaussian noise). The above can be explained in the sense that due to GPR-MANH is based on the assumption that

Table 1. UCI repository regression results. Bold: the method with the highest performance excluding the upper bound (target) classifier GPR-GOLD

(a) Homogeneous Gaussian noise					
Method	GPR-GOLD	GPR-Av	LMO	GPR-MA	GPR-MANH
	RMSE $\times$ 100				
Auto	35.40 ± 3.10	59.41 ± 7.03	70.30 ± 9.90	39.07 ± 9.41	$\mathbf{37.93 \pm 3.58}$
Concrete	40.05 ± 2.16	58.43 ± 1.95	71.89 ± 13.28	44.63 ± 2.39	$\mathbf{43.92 \pm 2.31}$
Housing	38.87 ± 5.90	56.85 ± 4.46	68.65 ± 11.48	38.96 ± 4.44	$\mathbf{38.92 \pm 4.50}$
Average	38.10	58.23	70.28	40.89	**40.26**
(b) Non-homogeneous Gaussian noise					
Method	GPR-GOLD	GPR-Av	LMO	GPR-MA	GPR-MANH
	RMSE $\times$ 100				
Auto	35.40 ± 3.10	54.49 ± 10.56	67.52 ± 10.91	36.82 ± 3.18	$\mathbf{36.42 \pm 3.14}$
Concrete	40.05 ± 2.16	54.46 ± 2.48	84.60 ± 8.67	43.96 ± 2.37	$\mathbf{42.28 \pm 2.13}$
Housing	38.87 ± 5.90	54.62 ± 3.86	74.72 ± 11.56	45.56 ± 5.81	$\mathbf{39.93 \pm 5.21}$
Average	38.10	54.52	75.61	42.11	**39.54**

the input space can be represented by a defined number of partitions, where each annotator exhibits a particular performance in each cluster. We highlight that the promising results of our approach are achieved based only on the responses from multiple annotators without considering any prior information.

5 Conclusion

In this paper, we presented a probabilistic framework based on Gaussian processes, termed GPR-MANH, to deal with regression problems in the presence of multiple annotators. Our approach relaxes the assumption that the performance of each annotator is homogeneous across the input space. GPR-MANH assume that the input space can be divided into K regions, where each annotator exhibit a particular level of expertise, which is coded by a variance $(\sigma_k^r)^2$. Then, the annotations are modeled as a version of the gold standard corrupted by additive and non-homogeneous Gaussian noise with zero mean and variance $(\sigma_k^r)^2$. Furthermore, we tested our approach using synthetic datasets from the UCI repository and simulate the annotations from multiple annotators following two different models (see Sect. 3). The results show that the proposed method can be used to perform regression problems in the context of multiple labelers with different levels of expertise. In fact, in most cases, our approach achieves better results when compared to different state-of-the-art techniques [1,9].

Finally, note that GPR-MANH loosens the assumption that the performance of the annotators only depends on the ground truth labels. As future work, this could be taken a step further by modeling the performance of the annotators as a function of the gold standard and the input samples through a Heteroscedastic Gaussian processes approach. Also, our method assumes independence between the opinions of the annotators; though it is suitable to consider that the labelers make their decisions independently, it is not true that these opinions are independent, due to there are possible correlations between the expert views. Accordingly, we expect to relax this assumption by using a probabilistic framework that allows to code the inter-annotator dependencies.

Acknowledgments. This work was funded by Colciencias under the project with code: 1110-744-55958. J. Gil González is funded by the program "Doctorados Nacionales - Convocatoria 785 de 2017". A. Orozco was partially funded by Maestría en ingeniería eléctrica from the Universidad Tecnológica de Pereira.

References

1. Groot, P., Birlutiu, A., Heskes, T.: Learning from multiple annotators with Gaussian processes. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds.) ICANN 2011. LNCS, vol. 6792, pp. 159–164. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21738-8_21
2. Wolley, C., Quafafou, M.: Learning from multiple annotators: when data is hard and annotators are unreliable. In: 2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW), pp. 514–521. IEEE (2012)

3. Mozetič, I., Grčar, M., Smailović, J.: Multilingual Twitter sentiment classification: the role of human annotators. PloS One **11**(5), e0155036 (2016)
4. Rodrigues, F., Lourenco, M., Ribeiro, B., Pereira, F.C.: Learning supervised topic models for classification and regression from crowds. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2409–2422 (2017)
5. González, J.G., Álvarez, M.A., Orozco, Á.A.: Automatic assessment of voice quality in the context of multiple annotations. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 6236–6239. IEEE (2015)
6. Raykar, V.C., et al.: Learning from crowds. J. Mach. Learn. Res. **11**, 1297–1322 (2010)
7. Rodrigues, F., Pereira, F.C., Ribeiro, B.: Gaussian process classification and active learning with multiple annotators. In: ICML, pp. 433–441 (2014)
8. Yan, Y., Rosales, R., Fung, G., Subramanian, R., Dy, J.: Learning from multiple annotators with varying expertise. Mach. Learn. **95**(3), 291–327 (2014)
9. Xiao, H., Xiao, H., Eckert, C.: Learning from multiple observers with unknown expertise. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds.) PAKDD 2013. LNCS (LNAI), vol. 7818, pp. 595–606. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37453-1_49
10. Bishop, C.M.: Pattern recognition. Mach. Learn. **128**, 1–58 (2006)
11. Rasmussen, C.E.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)
12. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**(5814), 972–976 (2007)