MOCA: An Unsupervised Algorithm for Optimal Aggregation of Challenge Submissions

Robert Vogel
IBM Thomas J. Watson Research Center
1101 Ketchawan Rd
Yorktown Heights, NY 10598

Mehmet Eren Ahsen
Gies College of Business
University of Illinois at Urbana-Champaign
Champaign, IL 61820
ahsen@illinois.edu

Gustavo A. Stolovitzky
IBM Thomas J. Watson Research Center
1101 Ketchawan Rd
Yorktown Heights, NY 10598
gustavo@us.ibm.com

The past decade has witnessed a significant increase in the diversity of machine learning methods as well as availability of data. These advances provide a fertile ground for crowd-sourced data challenges. An example of such data competitions in biomedicine is the DREAM (Dialogue for Reverse Engineering Assessment and Methods) Challenges [Saez-Rodriguez et al., 2016]. In the DREAM Challenges, organizers share data about an open biomedical problem, crowd-source its solution, and score the submissions against a blind gold standard.

Throughout numerous DREAM challenges we have consistently observed that by aggregating the submitted predictions we can get a more robust solution [Saez-Rodriguez et al., 2016]. The simplest way to combine ranked predictions to a classification problem is the "Wisdom of Crowds" (WOC) ensemble method, which averages the ranks assigned to each sample by the individual predictions thus creating a new ranked list of samples. Even though this simple strategy provides robust solutions, it is not optimal. In fact, it assigns the same weight to both weakly predictive and good classifiers. In this work we propose MOCA (Method for Optimal Classification by Aggregation), an ensemble algorithm that optimally combines predictions from multiple binary classification algorithms in an unsupervised setting. The unsupervised nature of MOCA is justified by the paucity of data in some important domains of biomedical research such as malaria, in which crowdsourcing can greatly help accelerate solutions to pressing global health problems [Davis et al., 2019].

Typical machine learning classification algorithms assign a score to each sample which can be used as a measure of the confidence that such sample belongs to the positive class. Prior research has shown that a proper calibration of classifier confidence scores improves the performance of ensemble algorithms [Whalen and Pandey, 2013]. In our unsupervised setting we will use the rank transformation as a calibration tool; hence, each method k ranks samples from the ones it most confidently assigns to the positive class (low ranks) to the ones it least confidently assigns to the positive class (high ranks). The input to MOCA, thus, is a matrix of ranked predictions $R \in \mathbb{R}^{N \times M}$, where r_{ik} is the rank assigned to sample $1 \leq k \leq N$ by method $1 \leq i \leq M$. Given the rank prediction matrix R, MOCA finds an optimal linear combination of individual methods of the following form:

$$s_k = \sum_{i=1}^{M} w_i r_{ik}, \tag{1}$$

where s_k is the MOCA score of the k^{th} sample, w_i is the coefficient of the i^{th} base classifier, and r_{ik} is the rank prediction of sample k by classifier i. Note that r_{ik} is known to us; our task, then, is to determine the optimal weights w_i. To do that we first define the conditional mean and variance of

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

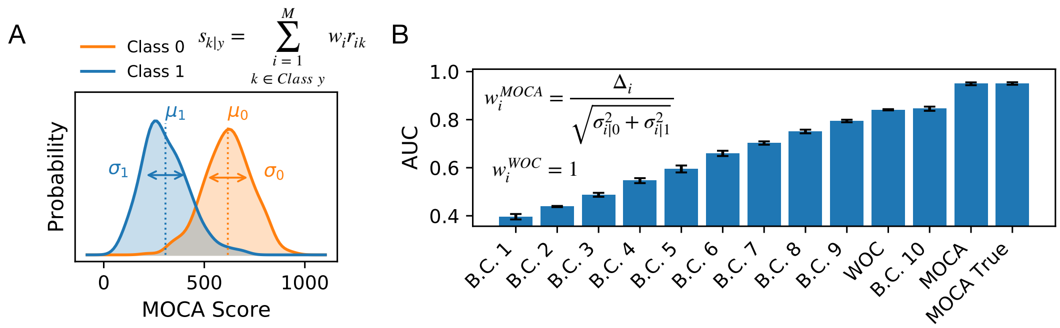

Figure 1: The predictions of $M = 10$ base classifiers (B.C. 1,...,10) with evenly spaced $0.4 \leq$AUC≤ 0.85, were generated for $N = 2500$ samples, of which $N_1 = 1000$ were from the positive class. The negative class scores were sampled from $\mathcal{N}(0, 1)$ while positive class samples from $\mathcal{N}(\mu_G, 1)$, where μ_G is a function of AUC [Marzban, 2004]. (A) The probability density of MOCA scores conditioned by the true class labels and (B) the mean AUC $\pm$ S.E.M. from 5 fold cross-validation.

MOCA scores:

$$\mu_0 = \frac{1}{N_0} \sum_{\{k|y_k=0\}} s_k, \quad \mu_1 = \frac{1}{N_1} \sum_{\{k|y_k=1\}} s_k \tag{2}$$

$$\sigma_0^2 = \frac{1}{N_0} \sum_{\{k|y_k=0\}} (s_k - \mu_0)^2, \quad \sigma_1^2 = \frac{1}{N_1} \sum_{\{k|y_k=1\}} (s_k - \mu_1)^2, \tag{3}$$

where y_k denotes the label associated with sample k, and N_0 and N_1 are the number of positive and negative samples, all unknown to us. We want to choose weights w_i such that the separation of distribution of MOCA scores is maximized between the two classes (See Figure 1A). Mathematically, we want to maximize the following quantity:

$$D := \frac{\mu_0 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_0^2}}. \tag{4}$$

In this work, we show that the MOCA weights maximizing (4) have the following closed form:

$$\mathbf{w}^{\text{MOCA}} = \frac{\mathbf{C}^{-1}\,\boldsymbol{\Delta}}{\left\|\mathbf{C}^{-1}\,\boldsymbol{\Delta}\right\|}, \tag{5}$$

where where $\boldsymbol{\Delta}$ is the $M \times 1$ vector with $\Delta_i = \mu_{i|0} - \mu_{i|1}$ where for each method i, $\mu_{i|0} = \frac{1}{N_0} \sum_{\{k|y_k=0\}} r_{ik}$, $\mu_{i|1} = \frac{1}{N_1} \sum_{\{k|y_k=0\}} r_{ik}$, and $\mathbf{C} = \mathbf{C}_0 + \mathbf{C}_1 \in \mathbb{R}^{M \times M}$ is the sum of the conditional covariance matrices $\mathbf{C}_y$ where each element $C_{ij|y}$ is the covariance of class conditioned ranks between methods i and j for $y = 0, 1$. Although we proved that the MOCA weights given in (5) are the unique optimizers of the function D defined in (4), calculation of both $\boldsymbol{\Delta}$ and $\mathbf{C}$ requires the knowledge of class labels which we assume is not observed. Under the assumption that the classifiers make conditionally independent predictions given the class, we have previously shown [Ahsen et al., 2018] how to estimate Δ in an unsupervised manner using the covariance matrix between classifier predictions. Here, we extend their result and prove a novel way to estimate $\mathbf{C}_0$ and $\mathbf{C}_1$, the conditional covariance matrices, in an unsupervised way using the third order covariance tensor between classifier predictions.

We demonstrate the performance of MOCA using simulation data. Figure 1A shows the schematic representation of the quantity D that is maximized by the MOCA weights. Figure 1B shows the AUC of M=10 classifiers and of two different strategies for their aggregation: the WOC ensemble where each method is given the same weight, and the MOCA ensemble, where the weights are computed in an unsupervised way using Eq. 5. For comparison we also show the MOCA ensemble if we used the MOCA true weights (MOCA True), that is, the weights computed with full knowledge of labels rather than in an usupervised way. We can see that the MOCA ensemble substantially outperforms the best individual model and the WOC strategy, and is comparable with the MOCA True ensemble. The unsupervised nature of MOCA makes it attractive to many applications including transfer learning. MOCA can be a powerful tool to aggregate solutions submitted to crowd-sourced data challenges, to optimally aggregate the collective intelligence of the crowd.

References

Mehmet Eren Ahsen, Robert Vogel, and Gustavo Stolovitzky. Unsupervised evaluation and weighted aggregation of ranked predictions. *arXiv preprint arXiv:1802.04684*, 2018.

Sage Davis, Katrina Button-Simons, Taoufik Bensellak, Eren Mehmet Ahsen, Lisa Checkley, Gabriel J Foster, Xinzhuan Su, Ahmed Moussa, Darlington Mapiye, Sok Kean Khoo, et al. Leveraging crowdsourcing to accelerate global health solutions. *Nature biotechnology*, 37(8):848–850, 2019.

Caren Marzban. The roc curve and the area under it as performance measures. *Weather and Forecasting*, 19(6):1106–1114, 2004.

Julio Saez-Rodriguez, James C Costello, Stephen H Friend, Michael R Kellen, Lara Mangravite, Pablo Meyer, Thea Norman, and Gustavo Stolovitzky. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nature Reviews Genetics*, 17(8):470, 2016.

Sean Whalen and Gaurav Pandey. A comparative analysis of ensemble classifiers: case studies in genomics. In *2013 IEEE 13th International Conference on Data Mining*, pages 807–816. IEEE, 2013.