Optimism in Active Learning with Gaussian Processes

Timothé Collet, Olivier Pietquin

Optimism in Active Learning with Gaussian Processes

Timothé Collet [†] and Olivier Pietquin [‡]

[†] CentraleSupélec, MaLIS Research group, France
GeorgiaTech-CNRS UMI 2958, France
timothe.collet@centralesupelec.fr
[‡] Univ. Lille - CRIStAL Lab (UMR 9189), France
IUF (Institut Universitaire de France)
olivier.pietquin@univ-lille1.fr

Abstract. In the context of Active Learning for classification, the classification error depends on the joint distribution of samples and their labels which is initially unknown. The minimization of this error requires estimating this distribution. Online estimation of this distribution involves a trade-off between exploration and exploitation. This is a common problem in machine learning for which multi-armed bandit theory, building upon Optimism in the Face of Uncertainty, has been proven very efficient these last years. We introduce two novel algorithms that use Optimism in the Face of Uncertainty along with Gaussian Processes for the Active Learning problem. The evaluation lead on real world datasets shows that these new algorithms compare positively to state-of-the-art methods.

Keywords: Active Learning, Classifiaction, Multi-Armed Bandits

1 Introduction

Classification is a supervised learning framework consisting in the association of instances to labels, existing in finite number. It uses a set of instances already associated to labels called the training set. In order to get a high prediction accuracy, the training set should contain a high number of instances. Recently, it has become easy to collect, store and process large data sets. However, the association of each instance of the training set to labels requires the manual annotation by an expert. This task is time consuming, and may involve other costs.

In Active Learning, the goal is to minimize the number of requests to the expert needed to achieve a targeted performance of the classifier. Equivalently, to maximize its performance under a fixed budget of requests. Hence, an Active Learning algorithm dynamically constructs the training set by sequentially deciding which instance to present to the expert. The instance is chosen considering all the previously received labels, such that its inclusion in the training set lead to the best performance of the classifier. We work under the pool-based

sampling scheme in which a set of unlabeled instances is available. The Active Learning algorithm successively selects instances from the pool to be labeled.

Among the many ways that exist to formulate this problem, which are reviewed in [8], this work focus on the Error Minimization framework. In this framework, the selection criterion is derived from the true risk, or misclassification rate. It is thus directly linked to the (mostly used) measure of performance for classifiers. We distinguish two strategies in this framework: the function to be minimized is either the maximum or the average misclassification rate among the instance space. In the first case, the strategy is to estimate the true risk related to each unlabeled instance and to sample the one for which it is maximum. In the second case, the strategy is to simulate the sample of each instance and to effectively sample the one resulting in the maximum decrease of the true risk. This second strategy has the advantage of representativeness, which means that an instance with a lot of unlabeled data around it is preferred to an isolated one.

The true risk can not be known perfectly, since it requires the true label distribution which need to be learned. But it can be estimated, along with a measure of uncertainty, using the labels received so far. Among the methods that exist in Error Minimization, the difference lies in how to manage the uncertainty about the true risk. In [9], the true risk is directly replaced by its estimation. In [4] and [5], the binary loss is expected over the possible values of the true conditional density of the class label. In [3], a min-max approach is used to ensure a minimum decrease of the risk.

In this paper, we propose a new solution to manage the uncertainty in the Error Minimization problem based on the Optimism in the Face of Uncertainty approach. Indeed, querying an instance may be used in order to increase the performance of the classifier directly, but also to improve future decisions. This is thus a case of exploration/exploitation dilemma for which the Optimism in the Face of Uncertainty is a well-renown approach, particularly for finite budget problems. The method is thus to establish first a probability bound on the value of the ideal criterion. Then to select the instance for which it is maximum.

Optimism in the Face of Uncertainty has been successfully used for Active Learning. In [2], the authors focused on the reduction of the version space. In [1], the instance space is partitioned and an allocation strategy is defined.

We apply our solution to Gaussian Processes. Indeed, they provide a simple Bayesian learning framework which outputs a probability distribution on the conditional density of the class labels. It is thus well fitted for Active Learning, and particularly for Optimistic methods.

We briefly review Gaussian Processes in Section 2. Then, we present our proposed algorithms for both strategies in Sections 3 and 4. We present the experiments end their results in Sections 5 and 6.

2 Gaussian Processes

A stochastic process is a generalization of a probability distribution to functions, each input is associated with a probability distribution on the output. In the case

of Gaussian Processes (GP), this distribution is Normal and can therefore be characterized only by its mean and variance. It thus provides a simple Bayesian framework for the supervised learning problem.

Let us consider the instance space X and the label set Y. In binary classification, the label set is composed of two elements, here $Y = \{-1, 1\}$. In the case of noisy classification, the expert can be represented by a Bernoulli distribution:

$$P(y \in Y = 1 | x \in X) = \frac{\mu(x) + 1}{2},$$

where $\mu(x)$ is the mean label for instance x. We use a GP to estimate the values of $\mu(x)$ from $\mathcal{S}$.

Let us denote $\mathcal{D}_p$ the set containing all the instances from the pool, it is split into $\mathcal{D}_l$ and $\mathcal{D}_u$, respectively the sets of labeled and unlabeled examples of size n_l and n_u. Suppose that we are given a training set $\mathcal{S} = \{(x_i, y_i)\}_{i \in [\![1, n_l]\!]}$ containing all the labeled instances and the associated label given by the expert.

A GP make use of a kernel function $k(.,.)$ to specify the covariance between outputs. Let $K_{l,l} = [k(x_i, x_j)]_{n_l \times n_l}$ be the covariance matrix of labeled instances. Let $K_{x,l} = [k(x, x_i)]_{x_i \in \mathcal{D}_l}$ be the covariance of an instance $x \in X$ and the labeled instances.

Conditioning the joint Gaussian distribution of $\mu(x)$ on the observed labels gives the following joint posterior distribution:

$$\mu(x) | \mathcal{S} \sim \mathcal{N}(m_{\mathcal{S}}(x), \sigma^2_{\mathcal{S}}(x)).$$

with $m_{\mathcal{S}}(x) = K_{x,l}(K_{l,l} + \sigma^2 I)^{-1} \mathbf{y}_l$ and $\sigma^2_{\mathcal{S}}(x) = k(x,x) - K_{x,l}(K_{l,l} + \sigma^2 I)^{-1} {K_{x,l}}^\top$.

The GP classifier predicts labels according to the sign of the posterior distribution's mean.

$$l(x) = sign(m_{\mathcal{S}}(x)) \tag{1}$$

Note that this mean is the same as the Regularized Least Square Regression (RLSR). The interest of GP is in Active Learning, where the confidence measure is able to tell in which region of the instance space the model has to be refined.

3 Local Risk Minimization

In this section, we propose an algorithm based on Optimism in the Face of Uncertainty that minimizes the maximum local risk of the classifier. We first define the loss function then derive an criterion that tend to minimize it.

Suppose that we are given a budget n. At each time step $t \in [\![1, n]\!]$, the Active Learning algorithm selects an unlabeled instance $x_t \in \mathcal{D}_u$, submits it to the expert, receives a label y_t and adds the pair to the training set.

At any time step t, the current training set $\mathcal{S}_t$ is used to train the GP. This one is then used to predict the label $l_t(x)$ for any instance $x \in X$ from Eq. (1). The expected prediction error, or local risk, is the probability that the expert would give a different label:

$$r_t(x) = \begin{cases} \frac{1+\mu(x)}{2} \text{ if } l_t(x) = -1 \\ \frac{1-\mu(x)}{2} \text{ if } l_t(x) = +1 \end{cases} . \tag{2}$$

We want to minimize the maximum risk one would encounter by presenting an unknown instance $x \in X$. We therefore define the following loss:

$$L_t = \arg\max_{x \in X} r_t(x). \tag{3}$$

In order to minimize this kind of loss, the solution is to select the instance for which the local risk is maximum.

$$x_{t+1} = \arg\max_{x \in \mathcal{D}_u} r_t(x).$$

Indeed, sampling an instance will necessarily lower the risk at its location, sampling where the maximum risk is attained guarantees to decrease this maximum:

However, this loss can not be used as such. Indeed, this suppose to know the true value of $\mu(x)$ at least for every unlabeled instances, which is unrealistic. If it was known, the prediction would be direct. We propose to use Optimism in the Face of Uncertainty in order to approach this solution knowing only a distribution of $\mu(x)$, given by the GP. In this paradigm, a confidence bound on the ideal criterion is establish relatively to a probability δ, then its upper bound is used as a heuristic. We show how to use the distribution of $\mu(x)$ to derive the distribution of the local risk.

Suppose that the label $l_t(x) = -1$ is predicted. Then,

$$\mathbb{P}(r_t(x) \geq \epsilon_t(x)) = \mathbb{P}(\mu(x) \geq 2\epsilon_t(x) - 1).$$

The GP trained with labeled instances outputs the following distribution:

$$\mu(x)|\mathcal{S}_t \sim \mathcal{N}(m_{\mathcal{S}_t}(x), \sigma^2_{\mathcal{S}_t}(x)).$$

Thus,

$$\mathbb{P}(r_t(x) \geq \epsilon_t(x) = \delta \iff 2\epsilon_t(x) - 1 = \Phi^{-1}(\delta, m_{\mathcal{S}_t}(x), \sigma^2_{\mathcal{S}_t}(x)),$$

with Φ the cumulative distribution function of the Normal distribution.

This applies symmetrically for $l_t(x) = 1$. Let us remind that $l_t(x) = sign(m_{\mathcal{S}_t}(x))$ from Eq (1), then:

$$\epsilon_t(x) = \frac{1 + \Phi^{-1}(\delta, |m_{\mathcal{S}_t}(x)|, \sigma^2_{\mathcal{S}_t}(x))}{2}. \tag{4}$$

Thus, ϵ_t bounds r_t with probability $1 - \delta$. Following an optimistic approach our algorithm selects at each time step the unlabeled instance for which ϵ_t is maximum:

$$x_{t+1} = \arg\max_{\mathcal{D}_u} \epsilon_t(x_p). \tag{5}$$

Minimizing the local risk is a solution that has shown good results in state-of-the-art. However, it does not consider the density of instances around the considered instances. Indeed, the criterion is computed using only their own prediction from the GP. In the next section, we change the loss to consider this.

4 Global Risk Minimization

In this section, we propose an algorithm based on Optimism in the Face of Uncertainty that minimizes the global risk of the classifier. The main improvement of this new algorithm compared to the one of the previous section is that it considers the representativeness (as defined in [**?**]) of an instance. We first the loss function being used and then show a criterion that tend to minimize it.

We now consider the expected error one would encounter by presenting an instance $x \in X$ drawn from the instance distribution which is the global risk of the classifier:

$$R_t(\mathcal{S}_t) = \int_X r_t(x) dP(x).$$

It is impractical because of the integral and because we do not known exactly the instance distribution. However, we have access to a pool instances following this distribution. We can thus estimate the global risk from it. We therefore define the loss:

$$\tilde{R}_t(\mathcal{S}_t) = \sum_{x \in \mathcal{D}_p} r_t(x).$$

Defining a strategy that optimally minimizes this loss is NP-hard. However, it is common to make a myopic approximation. Thus the ideal strategy is to simulate the sampling of all the instances in the pool of unlabeled instances. Then, to effectively sample the one which results in the highest decrease of the risk:

$$x_{t+1} = \arg\max_{x \in \mathcal{D}_u} \Delta_{x_s} \tilde{R}_t(\mathcal{S}_t), \tag{6}$$

where

$$\Delta_{x_s} R_t(\mathcal{S}_t) = \tilde{R}_t(\mathcal{S}_t) - \tilde{R}_t(\mathcal{S}_t \cup \{x_s, y_s\}),$$

where y_s is the true label of x_s.

This suppose to know the true values of $\mu(x)$ and y_s, which is unrealistic. We propose to use the Optimism in the Face of Uncertainty approach in order sample as close as possible to the ideal sampling but with only a distribution of $\mu(x)$, given by the GP. We show how to compute a confidence bound on the criterion from Eq. (6) relatively to a probability δ.

Let us consider the simulation of the sample of $x_s \in \mathcal{D}_u$ and that the label given by the expert is y_s. Let $l_t^+(x) = sign(m_{\mathcal{S}_t \cup \{x_s, y_s\}}(x))$. Then,

$$\forall x \in \mathcal{D}_p, \quad r_t(x) = \mathbb{1}_{l_t(x) \neq sign(\mu(x))} - \mathbb{1}_{l_t^+(x) \neq sign(\mu(x))},$$

Thus,

$$\forall x \in \mathcal{D}_p, \quad r_t(x) = \begin{cases} \mathbb{1}_{l_t(x) \neq -1} - \mathbb{1}_{l_t^+(x) \neq -1} \text{ with prob. } \mathbb{P}(\mu(x) < 0 | \mathcal{S}_t \cup \{x_s, y_s\}) \\ \mathbb{1}_{l_t(x) \neq 1} - \mathbb{1}_{l_t^+(x) \neq 1} \text{ with prob. } \mathbb{P}(\mu(x) \geq 0 | \mathcal{S}_t \cup \{x_s, y_s\}). \end{cases}$$

We can then combine the cases for every instance in the pool to deduct the probability: $\mathbb{P}(\Delta_{x_s} R_t(\mathcal{S}_t) | \mathcal{S}_t \cup \{x_s, y_s\})$. We can also combine the cases for y_s,

given that $\mathbb{P}(y_s = +1) = \mathbb{P}(\mu(x_s) \geq 0|\mathcal{S}_t)$, to deduct: $\mathbb{P}(\Delta_{x_s} R_t(\mathcal{S}_t)|\mathcal{S}_t)$. We then compute the probability bound e_t relatively to a fixed probability δ as:

$$\mathbb{P}(\Delta_{x_s} R_t(\mathcal{S}_t) \leq e_t(x_s)|\mathcal{S}_t) = 1 - \delta. \tag{7}$$

Following an optimistic approach, our algorithm selects at each time step the unlabeled instance for which e_t is maximum:

$$x_{t+1} = \underset{x_s \in \mathcal{D}_u}{\arg\max}\, e_t(x_s). \tag{8}$$

5 Experiments

We evaluated our algorithms on several datasets from the UCI repository [7]. Each experiment consisted in a series of 1,000 runs. For each of these runs, the dataset was randomly divided in two equal parts: one was used as the pool of instances in which the algorithms were allowed to pick, the other one was used as a test set, hidden from the algorithm. Our algorithm is initialized with one labeled instance randomly drawn from the pool. At each time step, the accuracy of the classifier is recorded as the proportion of well classified instances on the test set. The global performance displayed on figures results from averaging the accuracy of every run.

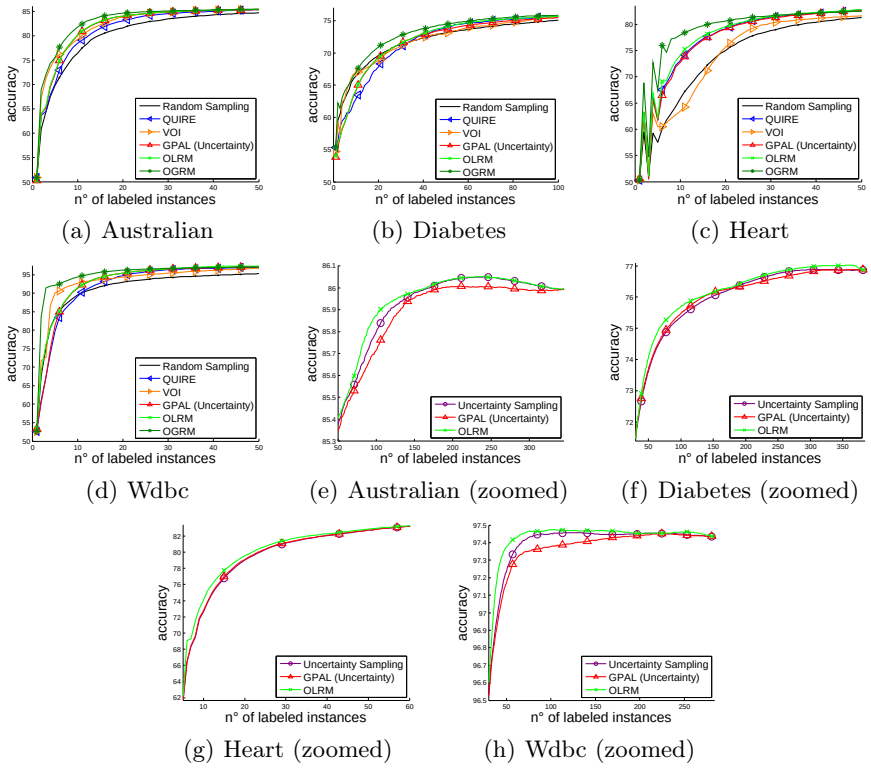
We compare OLRM and OGRM with the following methods: (1) Random Sampling: each instance to be labeled is randomly drawn from the pool, (2) Uncertainty Sampling (US) [6] samples instance closest to the boundary, (3) QUIRE [3] min-max strategy for minimization of the global risk, (4) VOI [5] minimization of the expected global risk, and (5) GPAL (Uncertainty) [4] minimization of the expected local risk.

All these algorithms use the GP or RLSR classifier. Indeed, different classifier have different intrinsic performances. Thus, we think it is not meaningful to compare Active Learning algorithms which are based on different classifiers. Here, the Gaussian kernel is used. The optimal parameter σ has been found using grid-search.

6 Results

Fig. 1[a-d] displays the classification accuracy for each evaluated methods for with different budget of labeled instances. Notice that even though the marker only appears ten times on each curve, the accuracy is plotted for every integer number of labeled instances. As some curves are hard to differentiate, Fig. 1[e-h] shows a zoomed version of, respectively, Fig. 1[a-d], focusing on those methods.

The parameter δ in OLRM and OGRM controls the exploration/exploitation trade-off. In the case of OGRM, this parameter was tuned using a gridsearch. It appeared that any value lower than 0.5 lead to almost exactly the same performance, with a slight tendency for values closer to 0.5.



(a) Australian (b) Diabetes (c) Heart

(d) Wdbc (e) Australian (zoomed) (f) Diabetes (zoomed)

(g) Heart (zoomed) (h) Wdbc (zoomed)

Fig. 1. Evaluation of algorithms

In the case of OLRM, no value of the parameter had an outstanding performance. The optimal parameter always depended on the number of labeled instances considered. However, optimistic algorithm are often studied through a finite budget framework. This suppose that the budget is known from the beginning and the parameter may be chosen accordingly. Though, we do not provide any theoretical analysis to chose its value. The displayed curve for this algorithm shows the maximum performance OLRM would have for each budget if the parameter was chosen accordingly.

We can see on Fig. 1[a-d] that the three algorithms based on minimization of the local risk have very close performance. By looking at the zoomed version on Fig. 1[e-h], we see that they are not exactly equivalent. First, notice that US is strictly equivalent with OLRM with $\delta = 0.5$. While GPAL is designed as an improvement of US that considers the variance outputted by the GP, it is most of the time outperformed by it. We can also see that, for a majority budget, there exist a parameter of ORLM that can achieve better performance than US.

We observe on Fig. 1[a-d] that OGRM outperforms all the methods it is compared to. Using the global risk loss function inherently make the performance increase sooner. Indeed, it considers the density of instances, and has an a pri-

ori about where to sample with no observations. It is therefore surprising that QUIRE behaves worse than local risk methods. On Fig. 1[a-b], VOI starts with the same performance as OGRM but does not follow. This is because it considers the decrease of the estimation of the risk and not the estimation of the decrease of the risk.

7 Conclusion

In this paper, we introduce a new way to deal with uncertainty in the Error Minimization problem that uses Optimism in the Face of Uncertainty. We present two algorithms that work with two different losses. One can be seen as a generalization of Uncertainty Sampling, while the second extend the former to consider the instances distribution. Evaluations on real-world datasets were made that showed that both algorithms compared positively to state-of-the-art methods. Future work will focus on theoretical analysis and on the adaptation to the multi-class case. We will also consider improving the noise distribution being used to stick to the expert's distribution.

References

1. Collet, T., Pietquin, O.: Active learning for classification: An optimistic approach. In: Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), 2014 IEEE Symposium on. pp. 1–8. IEEE (2014)
2. Ganti, R., Gray, A.G.: Building bridges: Viewing active learning from the multi-armed bandit lens. arXiv preprint arXiv:1309.6830 (2013)
3. Huang, S.J., Jin, R., Zhou, Z.H.: Active learning by querying informative and representative examples. In: Advances in neural information processing systems. pp. 892–900 (2010)
4. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Active learning with gaussian processes for object categorization. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. pp. 1–8. IEEE (2007)
5. Kapoor, A., Horvitz, E., Basu, S.: Selective supervision: Guiding supervised learning with decision-theoretic active learning. In: IJCAI. vol. 7, pp. 877–882 (2007)
6. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 3–12. Springer-Verlag New York, Inc. (1994)
7. Lichman, M.: UCI machine learning repository (2013), `http://archive.ics.uci.edu/ml`
8. Settles, B.: Active learning literature survey. University of Wisconsin, Madison 52(55-66), 11 (2010)
9. Zhu, X., Lafferty, J., Ghahramani, Z.: Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In: ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining. pp. 58–65 (2003)