Nonparametric Online Machine Learning with Kernels

Khanh Nguyen
Centre for Pattern Recognition and Data Analytics, Deakin University, Australia
nkhanh@deakin.edu.au

Abstract

Max-margin and *kernel methods* are dominant approaches to solve many tasks in machine learning. However, the paramount question is how to solve model selection problem in these methods. It becomes urgent in online learning context. Grid search is a common approach, but it turns out to be highly problematic in real-world applications. Our approach is to view max-margin and kernel methods under a Bayesian setting, then use Bayesian inference tools to learn model parameters and infer hyper-parameters in principle ways for both batch and online setting.

1 Introduction

Max-margin is a powerful principle to construct learning model with high generalization capacity. This principle can be applied straightforward for linear case. To capture data distribution with non-linear nature, the input data are transformed onto a higher-or-infinite dimensional feature space via a kernel function, then fit to a linear model. The linear decision boundary in the feature space is a set of non-linear contours in the input space which is sufficient to describe data variety. Methods using this principle are named kernel methods. These methods depend solely on the kernel function which describes the similarity between two data instances, hence they can deal with different kinds of data, such as data with variable length.

Nevertheless, existing max-margin and kernel methods have no effective way to solve model selection problem, i.e, to find the optimal hyper-parameters. In practice, grid search is a common approach to tune hyper-parameters for a given dataset. However, the grid search suffers from two key drawbacks. First, the number of trials grows exponentially with the number of hyper-parameters. Although one can parallelize the grid search, it takes an expensive computational resource. Second, the values of hyper-parameters can be continuous and unbounded whilst the grid contains discrete values only, hence there is no guarantee that the result is optimal.

Model selection problem becomes urgent in online learning context where data come continuously, sequentially and evolves rapidly. Because the training set changes all the time, the optimal hyper-parameters at different moments might be totally different. Therefore, the learning system must keep track all models corresponding with all possible values of hyper-parameters to obtain the optimal predictive performance.

Bayesian inference is a powerful tool to model a particular learning problem via a graphical model. Model parameters and hyper-parameters are represented as latent nodes in graphical model whilst each data instance in the training set is treated as an observed node. From this view, we can apply Bayesian inference strategies, such as Markov chain Monte Carlo (MCMC), variational inference, to learn the most likely model parameters and infer hyper-parameters.

In this research, we want to conjoin body of two mature theories: kernel-based methods and Bayesian modeling to advance learning methods based on kernels and max-margin principle. Our research objective is to develop new learning methods at the intersection of kernel-based methods and Bayesian modeling. Our aims are to solve model selection problem for kernels and max-margin principle. Our methods are able to avoid grid search completely, thus reduce computational resources significantly. The remarkable point is that our methods can work with large-scale dataset and get ready for online learning setting.

Leveraging Bayesian approach with kernel-based methods has been investigated [Zhu *et al.*, 2011; Wang and Zhu, 2014]. However, they did not address model selection problem. To our best knowledge, our research is the first study that uses Bayesian inference to solve model selection in kernel methods and addresses model selection in online learning context. For the rest of paper, we present our first achievements and directions for the remaining work.

2 Multiple Kernel Learning Approach

In single kernel learning, the kernel function is usually a RBF kernel as follows

$$\kappa\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left\{-\gamma \left\|\mathbf{x} - \mathbf{x}'\right\|^2\right\}$$

However, there is a wide spectrum of linear or nonlinear kernel functions to choose and each kernel function has its own parameters to tune. A common approach is to run a grid search over sets of parameters to obtain the optimal one, but it is computationally expensive. A notable approach to relax the grid search is to use multiple kernel learning (MKL) [Gönen and Alpaydın, 2011]. In MKL approach, rather than using a single kernel, one prefers combining a wide spectrum of kernels into a linear weighted sum of kernels as follows

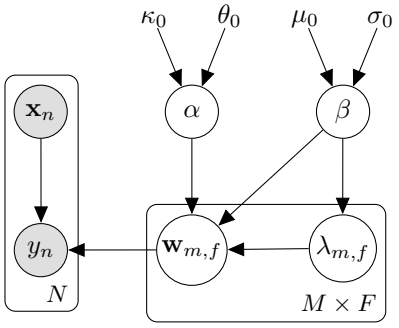

Figure 1: Graphical model of MKL approach.

$$\kappa(\mathbf{x}, \mathbf{x}') = \pi_1 \kappa_1(\mathbf{x}, \mathbf{x}') + \ldots + \pi_F \kappa_F(\mathbf{x}, \mathbf{x}')$$

By this way, the expressiveness capacity is increased and the best combination is automatically discovered by solving an optimization problem [Orabona and Jie, 2011]:

$$\min_{\mathbf{W}} \left(\frac{\alpha}{2} \|\mathbf{W}\|_{2,2}^2 + \beta \|\mathbf{W}\|_{2,1} + \sum_{n=1}^{N} l(\mathbf{W}; \mathbf{x}_n, y_n) \right) \quad (1)$$

where $\mathbf{W} = [\mathbf{w}_{1,1}, \ldots, \mathbf{w}_{1,F} \mid \ldots \mid \mathbf{w}_{M,1}, \ldots, \mathbf{w}_{M,F}]$, $\mathbf{w}_{m,f}$ is the hyperplane associated with m-th class in f-th feature space and $l(\mathbf{W}; \mathbf{x}_n, y_n)$ is a loss function.

However, this approach still requires a grid search to tune hyper-parameters α and β. Our solution is to view MKL under Bayesian view whose a (maximum a posteriori) MAP estimation reduces exactly to the optimization problem (1). Then, we construct a graphical model for MKL approach and do posterior inference. Because the posterior inference is intractable, we employ data augmentation technique [Polson and Scott, 2011] by coupling each $\mathbf{w}_{m,f}$ with an auxiliary variable $\lambda_{m,f}$. Finally, we have a graphical model as presented in Figure 1. As the result, we also avoid the group norm $\mathcal{L}_{2,1}$ that makes the optimization problem of MKL complicated. Then, it allows us to scale up and deal with online learning by applying Stochastic Gradient Descent framework. We validate our method on several benchmark datasets in both batch and online settings. The experimental results show that our proposed method can learn hyper-parameters in a principled way to eliminate the expensive grid search while gaining a significant computational speedup comparing with the state-of-the-art baselines [Nguyen *et al.*, 2016a].

3 Multi-Hyperplane Machine Approach

To address the scalability issue in kernel methods, multi-hyperplane machine approach is to learn a set of hyperplanes in the input space instead of learning in the feature space. The optimization problem is of the form

$$\min_{\mathbf{W}} \left(\frac{\alpha}{2} \|\mathbf{W}\|_{2,2}^2 + \sum_{n=1}^{N} l(\mathbf{W}; \mathbf{x}_n, y_n) \right)$$

where $l(\mathbf{W}; \mathbf{x}_n, y_n)$ is a loss function, $\mathbf{W} = [\mathbf{w}_{1,1}, \ldots \mathbf{w}_{1,K_1} \mid \ldots \mid \mathbf{w}_{M,1}, \ldots, \mathbf{w}_{M,K_M}]$ and K_m is the number of hyperplanes in m-th class. However, its solution is usually not sparse in terms of the number of hyperplanes for each class and the number of nonzero components in each hyperplane. This side effect can lead the model to be overfitted. [Wang *et al.*, 2011] addressed this issue by a heuristic pruning weight procedure, but it also hurt the predictive performance. In addition, a grid search is required for hyper-parameter tuning. For the first issue, we solved by using the group norm $\mathcal{L}_{2,1}$ coupled with $\mathcal{L}_{2,2}$ [Nguyen *et al.*, 2016b]. As we know in literature, by minimizing group norm $\mathcal{L}_{2,1}$, we can encourage the sparsity of the solution. However, this way also introduces a new hyper-parameter which also needs to be tuned. For the next step, we will solve the second issue by using Bayesian inference tools.

4 Conclusion

In this research, we have explored the use of Bayesian inference to solve model selection problem in kernel methods and max-margin principle for both batch and online settings. Our research can apply in many applications where they require to deal with large-scale datasets or streaming data and where they are incapable of using a time-consuming and expensive grid search for tuning hyper-parameters. In future work, we will address the model selection problem in semi-supervised learning where the training set contains not only labeled data but also unlabeled data.

References

[Gönen and Alpaydın, 2011] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

[Nguyen *et al.*, 2016a] Khanh Nguyen, Trung Le, Vu Nguyen, Tu Nguyen, and Dinh Phung. Multiple kernel learning with data augmentation. In *Proceedings of The 8th Asian Conference on Machine Learning*, pages 49–64, 2016.

[Nguyen *et al.*, 2016b] Khanh Nguyen, Trung Le, Vu Nguyen, and Dinh Phung. Sparse adaptive multi-hyperplane machine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 27–39. Springer, 2016.

[Orabona and Jie, 2011] Francesco Orabona and Luo Jie. Ultra-fast optimization algorithm for sparse multi kernel learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 249–256, 2011.

[Polson and Scott, 2011] Nicholas G. Polson and Steven L. Scott. Data augmentation for support vector machines. *Bayesian Anal.*, 6(1):1–23, 03 2011.

[Wang and Zhu, 2014] Yining Wang and Jun Zhu. Small-variance asymptotics for dirichlet process mixtures of svms. In *AAAI*, pages 2135–2141, 2014.

[Wang *et al.*, 2011] Z. Wang, N. Djuric, K. Crammer, and S. Vucetic. Trading representability for scalability: Adaptive multi-hyperplane machine for nonlinear classification. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 24–32, New York, NY, USA, 2011. ACM.

[Zhu *et al.*, 2011] Jun Zhu, Ning Chen, and Eric P Xing. Infinite svm: a dirichlet process mixture of large-margin kernel machines. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 617–624, 2011.