УДК 004.852

Алгоритм выбора классификатора и настройки его гиперпараметров на основе методов SMAC и Active Testing

Гуцол К. Д., Университет ИТМО, г. Санкт-Петербург
Научный руководитель – Муравьев С. Б., программист ФИТиП Университета ИТМО

Введение

Существует много алгоритмов классификации, учитывая их конфигурации, получаем огромное количество альтернатив. Выбрать лучшую пару алгоритм-конфигурация с помощью кросс-валидации вычислительно сложно, что делает невозможным применение этого решения в широком спектре задач. Появляется необходимость в оптимизации времени работы алгоритма. Существуют как алгоритмы оптимизации гиперпарарараметров и выбора алгоритма классификации, по отдельности, так и их комбинации, позволяющие решать объединенную задачу. В данной работе мы рассмотрим задачу выбора алгоритма классификации одновременно с подбором его гиперпараметров.

Постановка задачи

Имея множество алгоритмов классификации, информацию об их работе на предыдущих наборах данных и новый набор данных необходимо выбрать алгоритм, который будет лучшим относительно заданной метрики эффективности.

Базовые положения исследования

Были изучены существующие алгоритмы выбора алгоритма классификации и настройки его гиперпараметров. В частности, как Active Testing, SMAC и различные модификации этих алгоритмов.

Active Testing - алгоритм выбора алгоритма классификации. Он использует данные о результатах работы классификаторов на предыдущих наборах данных, чтобы подобрать наилучший классификатор для следующего набора данных.[6] Задачу выбора алгоритма классификации и настройки его гипермапаметров можно свести к задаче выбора алгоритма классификации, сгенерировав много пар (алгоритм, конфигурация), и выбирая уже из этих пар.[5] Этот подход характеризуется быстротой работы программы, но уступает в качестве решения модификациям SMAC-а, описаным ниже.

В рамках данной работы были рассмотрены модификации SMAC-а, которые используют данные о датасете, в частности, простые, статистические и теоретико-информационные характеристики, как объем тренировочных данных, количество отличительных признаков, классов, кроме того коэффициент асимметрии данных и их энтропию в процессе настройки гиперпараметров алгоритма.[1, 3, 4] Этот подход характеризуется высокой точностью решения и большой продолжительностью работы программы.

После исследования существующих методов решения был предложен новый, являющийся комбинацией двух описанных выше. В рамках этой работы данные о результатах работы различных классификаторов и их конфигураций на предыдущих

наборах данных также будет учитываться при настройке гиперпараметров, таким образом уменьшается время работы исходного алгоритма.

Список используемых источников

1. Hutter F., Hoos H. H., Leyton-Brown K. Sequential model-based optimization for general algorithm configuration //International Conference on Learning and Intelligent Optimization. – Springer, Berlin, Heidelberg, 2011. – С. 507-523.

2. Thornton C. et al. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms //Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2013. – С. 847-855.

3. Feurer M. et al. Efficient and robust automated machine learning //Advances in neural information processing systems. – 2015. – С. 2962-2970.

4. Feurer M. et al. Practical automated machine learning for the automl challenge 2018 //International Workshop on Automatic Machine Learning at ICML. – 2018

5. Leite R., Brazdil P., Vanschoren J. Selecting classification algorithms with active testing //International workshop on machine learning and data mining in pattern recognition. – Springer, Berlin, Heidelberg, 2012. – С. 117-131.

6. Leite R., Brazdil P. Active Testing Strategy to Predict the Best Classification Algorithm via Sampling and Metalearning //ECAI. – 2010. – С. 309-314.