

Обучение со слабой разметкой на основе регуляризации

Бериков В.Б.

ИМ СО РАН, Новосибирск, Россия
berikov@math.nsc.ru

В задачах обучения со слабой разметкой (weakly supervised learning [1]) аннотация данных может быть неполной, неточной или неоднозначной (например, из-за высокой стоимости, ошибок маркировки, привлечения группы разных разметчиков, большого числа наблюдений).

Для решения рассматриваемых задач разработаны различные подходы, основанные на адаптации моделей машинного обучения, в том числе вероятностные или ансамблевые методы, а также методы, основанные на предположениях о кластерной структуре данных и о регуляризации многообразия. Для построения векторных представлений объектов используется глубокое обучение.

Актуальной проблемой остается совершенствование указанных подходов: разработка методов, позволяющих за приемлемое время получать точные и надежные решения в условиях разнородности данных (различных форм их представления и источников получения) и недостатке обучающей информации; создание эффективных алгоритмов анализа данных большого объема. Существующие алгоритмы часто требуют трудоемких вычислений и большого объема памяти.

В данной работе будем рассматривать следующие способы задания слабой разметки и соответствующей постановки задачи.

а) Выборка имеет вид таблицы наблюдений за признаками объектов. Для каждого объекта a_i значение целевой переменной Y_i подчиняется некоторому известному распределению. Рассмотрим случай нормального распределения: $Y_i \sim \mathcal{N}(a_i, \sigma_i)$, параметры которого (либо их оценки) известны, например, оценены по данным краудсорсинга. Часть выборки неразмечена и используется как источник дополнительной информации. Требуется для каждого неразмеченного наблюдения построить решение в виде статистического распределения $F(x)$, т.е. на языке слабой разметки.

б) Рассматривается задача семантической сегментации изображений. Требуется выделить области двух классов, например патологические и здоровые участки на томографическом снимке. При этом обучающая разметка задается в виде масок, включающих области изображения интересующего класса. Неточность разметки проистекает из возможного низкого качества снимков, недостаточного опыта разметчиков, большого числа изображений, наличия шума в оборудовании и т.д.

В предлагаемом докладе рассматривается новый подход к решению задач обучения со слабой разметкой, основанный на различных вариантах регуляризации [2, 3, 4]. При этом для повышения эффективности решений используются такие методы, как латентное представление данных с помощью глубокого автоэнкодера; низкоранговые операции с матрицей сходства; сглаживание границ изображений. Приведены результаты применения разработанных алгоритмов и сравнения с рядом известных методов.

Работа проведена при поддержке РНФ, проект 24-21-00195.

Список литературы

1. Zhou Z.-H. A brief introduction to weakly supervised learning // National science review, 2018. V. 5(1). P. 44–53.
2. Kalmutskiy K. , Berikov V. Enhancing Stability of the Weakly Supervised Regression Algorithm Using Manifold Regularization and Fuzzy Clustering // Pattern Recognition and Image Analysis, 2025. V.35. No1. P.16–18.
3. Kalmutskiy K. , Berikov V. Clustering-Based Graph Neural Network in a Weakly Supervised Regression Problem // Lecture Notes in Computer Science, 2025. V.15681. P.335–347.
4. Mikhailapov D. , Tulupov A., Berikov V. Mask Correction in 3-D Tomography Brain Images for Weakly Supervised Segmentation of Acute Ischemic Stroke // 2024 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2024. IEEE. P. 796–801.