ОПТИМИЗАЦИЯ МЕТРИК В ЗАДАЧАХ С ЧАСТИЧНЫМ ОБУЧЕНИЕМ[1]

Г.В. Иофина, А.В. Минаев, Ю.С. Поляков, Ю.В. Максимов

Московский физико-технический институт, Московская область, Долгопрудный
e-mail: giofina@mail.ru

В докладе рассматривается задача машинного обучения с частичной информацией (semi-supervised learning problem, SSL problem). В классической постановке, входом задачи является l объектов с известным описанием и известной классификацией, а также u объектов с неизвестной классификацией, но известным описанием. При этом полагается, что $u \gg l$. Задача состоит в том, что бы восстановить классификацию неразмеченных объектов, среди которых могут быть объекты не входящие в заданную неразмеченную совокупность. При этом считается, что каждый класс описывается хотя бы одним размеченным объектом из обучающей совокупности.

К решению данной задачи существует множество подходов (см. например [4,5]). Как правило, алгоритмы решения SSL задач состоит из двух этапов. На первом этапе выделяются однородные области, в которых классификация может быть уверенно восстановлена на основе имеющейся размеченной совокупности. На втором этапе решается задача классификации с учетом "пополненной" совокупности.

В настоящей работе рассматриваются методы SSL обучения, основанные на пополнении обучающей совокупности на основе метода ближайших соседей. Классической проблемой непараметрических подходов этого типа (см. [5]) является экспоненциальная зависимость от размерности. Вклад авторов состоит в обобщении результатов работы [5] на случай многих классов. В работе также показано, что правильный выбор функции близости [2,3] с использованием идей, предложенных в [1], позволяет несколько снизить оценки минимального числа необходимых объектов от размерности в случае дискретных шкал.

ЛИТЕРАТУРА

1. N. Goyal, Y. Lifshits,H. Schütze. *Disorder inequality: a combinatorial approach to nearest neighbor search.* — WSDM '08 Proceedings of the International Conference on Web Search and Data Mining. — 2008. P. 25–32.
2. G.V. Iofina. *Optimal metrics in classification problems with ordered features and an arbitrary number of classes.* — Pattern Recognition and Image Analysis. — 2009. Vol. 19. Iss. 2. P. 284–288
3. G.V. Iofina. *A study of metrics in finite sets for application in classification and recognition problems.* — Computational Mathematics and Mathematical Physics. — 2010. Vol. 50. Iss. 3, P. 558–565
4. P. Rigollet. *Generalization Error Bounds in Semi-supervised Classification Under the Cluster Assumption.* — Journal of Machine Learning Research. — 2007. Vol. 8. P. 1369–1392.
5. R. Urner, S. Wulff, S. Ben-David. *PLAL: Cluster-based active learning.* — Journal of Machine Learning Research. Workshop and Conference Proceedings. — 2013. Vol. 30. P. 376–397.

[1]Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов №14-07-31241 мол_а и №14-07-31277 мол_а; а также Лаборатории структурных методов анализа данных в предсказательном моделировании, ФУПМ МФТИ, грант правительства РФ дог. 11.G34.31.0073