Иванов В.А., Шорников Е.В.
(Чебоксары, ООО «Релематика»)

ПОДХОД К ПОВЫШЕНИЮ ИНТЕРПРЕТИРУЕМОСТИ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧЕ ДЕТЕКТИРОВАНИЯ АНОМАЛИЙ

Одной из ключевых задач машинного обучения является задача детектирования (обнаружения) аномалий. Методы детектирования аномалий, как и многие методы машинного обучения, являются «черными ящиками». Однако понимание причин, лежащих в основе процесса прогнозирования весьма важно для оценки доверия и безопасности использования данных моделей.

В данной работе предложен способ объяснения моделей машинного обучения в задаче обнаружения аномалий. В его основе лежит методика объяснений LIME (Local Interpretable Model-agnostic Explanations) [1]. Методика LIME может объяснить предсказание любого классификатора путем создания локальной упрощенной линейной модели. В качестве алгоритма детектирования аномалий использован алгоритм Isolation Forest «изолирующий лес».

Методика LIME используется для объяснения моделей в задачах классификации и кластеризации [2], которые строят профиль обычных экземпляров, а затем идентифицирует экземпляры, которые не соответствуют нормальному профилю как аномалии. Алгоритм Isolation Forest использует совершенно иной принцип решения задач детектирования аномалий. Алгоритм основан на предположении, что аномалию намного легче изолировать, чем нормальный экземпляр. Isolation Forest строит ансамбль деревьев для набора данных, а затем находит аномалии как экземпляры, имеющие наиболее короткий средний путь в деревьях.

Для тестирования методики LIME для модели Isolation Forest использована обучающая выборка (рис. 1), на которой обозначены три точки, заведомо являющиеся аномальными. Следует отметить, что выбранные два признака представлены лишь для наглядности работы алгоритма. На рис. 2 представлен результат работы алгоритма для трех аномальных точек.

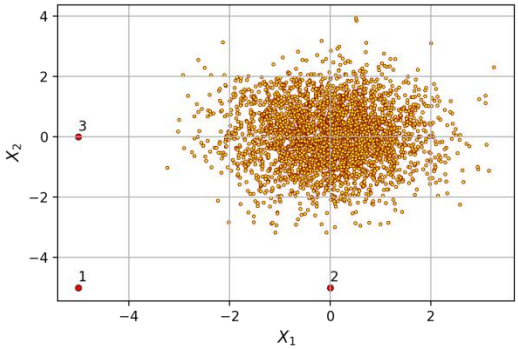

Рис. 1. Обучающая выборка

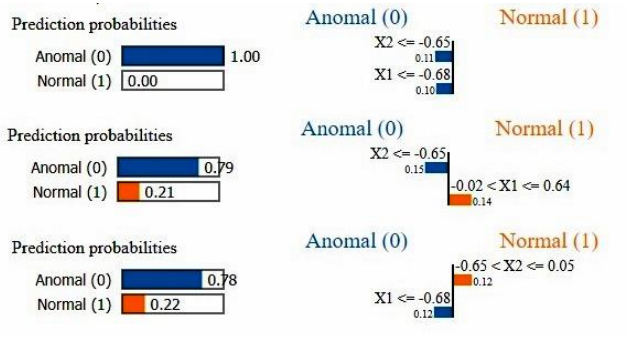

Рис. 2. Результат работы алгоритма

Алгоритмы машинного обучения для поиска аномалий могут применяться в релейной защите для обнаружения аномальных режимов в энергосистеме. Соответственно объяснение работы алгоритма детектирования аномалий имеет практическое значение в электроэнергетике.

Литература

1. *Ribeiro M. T., Singh S., Guestrin C.,* Why should I trust you?: Explaining the predictions of any classifier. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 2016.

2. *Z. He, X. Xu, S. Deng*. Discovering cluster-based local outliers. Pattern Recogn. Lett., 24(9-10):1641–1650, 2003.