Анализ применимости классических алгоритмов ml в практических задачах энергетики

Дементий Ю. А., ООО «Релематика», г. Чебоксары, Россия, dementiy.yu.a@gmail.com.

Петряшин А. Е., ООО «Релематика», г. Чебоксары, Россия, petryashin.a.e@mail.ru.

Петряшин И. Е., ООО «Релематика», г. Чебоксары, Россия, petryashin.i.e@mail.ru.

Аннотация: *одной из важных задач в энергетике и, в частности, релейной защите является разграничение режимов работы энергообъекта. В данной работе проанализированы основные подходы к решению этой задачи с использованием классических алгоритмов машинного обучения. Произведено сравнение следующих алгоритмов: метода ближайших соседей (KNN), дерева принятия решений и логистической регрессии. Также проведен анализ ограничений их использования. Изучена проблема влияния данных на правильность классификации режимов для простых и сложных моделей. Дана оценка применимости алгоритмов машинного обучения в рассмотренной задаче и сформулированы основные положения для дальнейших исследований.*

Ключевые слова: *машинное обучение, поиск границы области, классификация режимов работы энергообъекта.*

Введение

В данной работе рассматриваются возможности и ограничения классических алгоритмов машинного обучения (ML) для решения задачи разграничения режимов работы энергообъекта. На примере задачи построения релейной защиты анализируется возможность выполнения требований максимизации чувствительности с сохранением селективности. В качестве разграничиваемых режимов выступают отслеживаемые и альтернативные режимы (α- и β-режимы) объекта. Задача, а также ограничения на ее решение формулируются в терминологии машинного обучения.

Планирование эксперимента

Рассматривается задача построения алгоритма релейной защиты. Она решается на основе информации, накопленной в результате наблюдения за объектом (обучение по прецедентам) [1]. Под информацией понимаются точки наблюдаемого пространства, которые могут быть получены как с имитационной модели объекта (ИМО), так и с реального объекта.

В данном эксперименте источником информации является ИМО абстрактного объекта. Для наглядности выбрана ИМО, выходные величины которой принадлежат двухмерному наблюдаемому пространству с заранее известными границами областей отображаемых режимов. Внутри областей режимов точки располагаются случайным образом, что имитирует режимы работы реального объекта.

Алгоритм, построенный в результате решения поставленной задачи, должен соответствовать требованиям селективности и максимальной чувствительности, что эквивалентно одновременному выполнению условий (1) и (2)

$$recall_{\beta} = 1, \quad (1)$$

$$precision_{\alpha} \to \max. \quad (2)$$

На предмет соответствия этим условиям анализируются следующие алгоритмы классификации: метод k-ближайших соседей (KNN), дерево принятия решений (DTC) и линейные модели (LC).

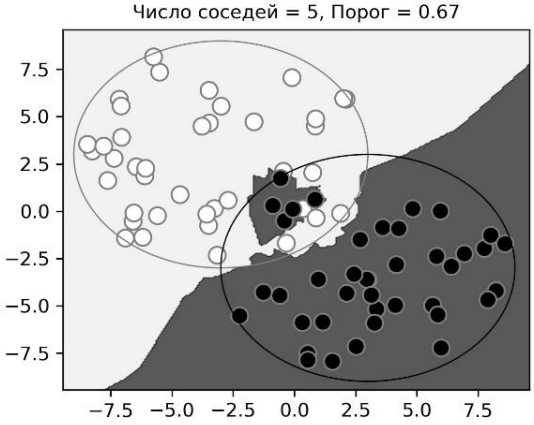
Выполнение условий (1) -(2) достигается путем изменения гиперпараметров модели. В качестве гиперпараметров, отвечающих за сложность моделей, выступают: число соседей – для KNN, глубина дерева – для DTC и размерность спрямляющего пространства – для LC, кроме того, для каждого алгоритма производится подбор весов классов (показатель важности класса в задаче оптимизации функции ошибки классификатора) или порога (минимальный уровень, после которого точка относится к классу β -режимов).

Применение алгоритмов

Рассматривается случай, когда α - и β -режимы ограничены в наблюдаемом пространстве окружностями, которые

пересекаются, причем каждый режим описывается 40 наблюдениями. Гиперпараметры алгоритмов подбираются таким образом, чтобы разделяющая кривая строилась с учетом условий (1) -(2).

На рисунках 1-3 представлены разделяющие кривые, построенные в результате работы алгоритмов, а также истинные границы режимов. Белым и черным показаны точки α - и β - режимов соответственно.



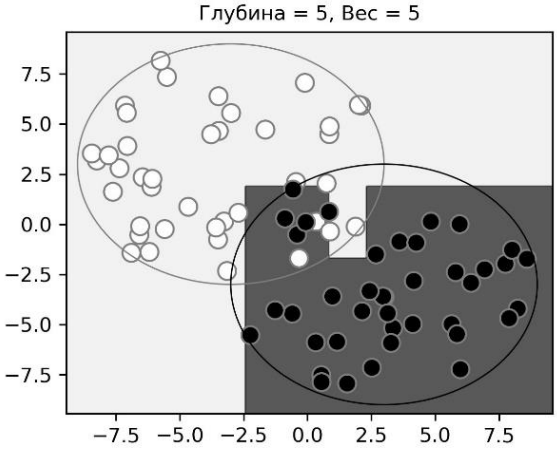


Рис.1. Разделяющая кривая, построенная алгоритмом KNN

Рис.2. Разделяющая кривая, построенная алгоритмом DTC

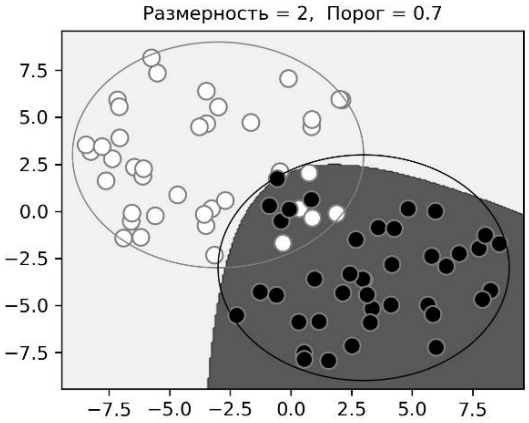


Рис.3. Разделяющая кривая, построенная LC

Для всех трех алгоритмов, несмотря на выполнение условия

(1), характерно классифицирование части области β-режимов, как области α- режимов. Это объясняется нехваткой данных вблизи границы области β-режимов – классифицируется неправильно та ее часть, в которой недостаточно точек, описывающих ее. Такое явление приводит к неверной классификации части области β-режимов, поэтому для построения алгоритма, разграничивающая линия которого соответствует истинной, необходимо получение дополнительной информации об областях с недостаточной плотностью точек, а также о граничных режимах.

Для алгоритма KNN характерно классифицирование части области β-режимов, в которой недостаточна плотность точек, как области α-режимов. Поэтому для правильного построения разделяющей кривой данным алгоритмом необходимо, чтобы плотность точек β-режимов была постоянной и большей, чем у α-режимов.

Из недостатков алгоритма DTC следует отметить форму разделяющей кривой - она состоит из участков, параллельных координатным осям. Из-за этого восстановление гладкой границы классов является затруднительным.

Алгоритм LC способен строить нелинейные гладкие разделяющие кривые. Повышение сложности алгоритма LC позволяет аппроксимировать границу классов с большей точностью, тем не менее чрезмерное увеличение сложности при нехватке данных может привести к переобучению.

Таким образом, ни один из представленных алгоритмов не позволяет решить задачу с выполнением условий (1) -(2). Для их выполнения необходимо увеличение количества информации до бесконечности, либо применение метода, позволяющего синтезировать информативные прецеденты для обучения. При необходимости построения разделяющей кривой по имеющимся данным необходимо подбирать гиперпараметры моделей таким образом, чтобы обобщающая способность модели не утрачивалась в результате чрезмерного повышения ее сложности. Тем не менее, влияние случайности данных не позволяет в общем случае утверждать возможность построения

алгоритма, соответствующего условиям (1) -(2).

Влияние сложности модели и случайности данных на границу режимов

С увеличением сложности моделей повышается их способность строить границы сложной формы, что при случайности данных может приводить к переобучению. Простые модели менее подвержены этому явлению, однако, они не всегда позволяют описывать границу достаточно точно. Поэтому в условиях недостатка данных необходимо соблюдение баланса между сложностью модели и ее способностью к обобщению данных (bias-variance tradeoff) [2].

Заключение

Главной проблемой, возникающей при обучении по прецедентам, полученным без учета их информативности, является невозможность построения классификатора, разделяющая линия которого соответствует истинной. Для приближения разделяющей границы к истинной необходимо увеличить число информативных данных, что при случайности их формирования оказывается труднодостижимым.

Простые модели менее чувствительны к недостатку информативных данных, однако они не обладают достаточной точностью, сложные модели при нехватке данных, а также их случайном характере, переобучаются и показывают плохие результаты. При переходе из двумерного пространства к многомерному проблема будет усугубляться, так как задача станет подвержена «проклятью размерности», поэтому требуемое количество информативных данных будет увеличиваться.

СПИСОК ЛИТЕРАТУРЫ

1. *Aurelien G.* Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2019, 1150 c.
2. Bias-Variance-Covariance Decomposition. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7687-1_932

Авторы:

Дементий Юрий Анатольевич, *ООО «Релематика», кандидат технических наук, руководитель группы, e-mail: dementiy.yu.a@gmail.com.*

Петряшин Александр Евгеньевич*, ООО «Релематика»/ЧГУ им. И.Н. Ульянова, инженер-исследователь, окончил в 2020 г. факультет Энергетики и Электротехники ЧГУ им. И.Н. Ульянова, получил степень бакалавра по направлению «Релейная защита и автоматизация электроэнергетических систем», e-mail: petryashin.a.e@mail.ru.*

Петряшин Илья Евгеньевич*, ООО «Релематика»/ЧГУ им. И.Н. Ульянова, техник-исследователь, студент факультета Энергетики и Электротехники ЧГУ им И.Н. Ульянова по направлению «Релейная защита и автоматизация электроэнергетических систем», e-mail petryashin.i.e@mail.ru.*