УДК 004

УГВУ ЧИНЕДУ КИНГСЛИ, студент группы BIT15
Университет Кади Айяд, КФУ
г. Марракеш

АЛГОРИТМ БЛИЖАЙШИХ СОСЕДЕЙ В ЗАДАЧЕ ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА ДАННЫХ

Общая формула алгоритма ближайшего соседа может быть представлена в виде (1).

$$w(i,x) = [i \leq 1] \quad (1)$$

В алгоритме используется вес i-го соседа для классификации объекта x; только первый сосед имеет вес, равный единице, а остальные соседи не учитываются [1, 2]. Объект в алгоритме относится к тому классу, к которому принадлежит его ближайший сосед. Недостатком алгоритма является тот факт, что по случайным причинам объект одного класса может попасть в сгусток объектов других классов. Тогда и данный объект, и соседние с ним объекты будут классифицироваться некорректно.

Такой тип алгоритма не может использоваться в аналитике точных данных [3], поскольку одна случайная аномалия существенно снизит всю обобщающую способность алгоритма искусственного интеллекта в целом. Однако надежность алгоритма ближайшего соседа можно существенно повысить, если исследовать не один объект «ближайшего соседа», а несколько объектов [4]. Причем в этой ситуации возникает параметр K:

$w(i,x) = [i \leq k]$ – метод k ближайших соседей.

Параметр k подбирается по критерию скользящего контроля (2):

$$LOO(k, X^l) = \sum_{i=1}^{l} [a(x_i, X^l \backslash \{x_i\}, k) \neq y_i] \rightarrow min_k \quad (2)$$

Из формулы (2) следует, что для каждого объекта x_i мы берем все объекты обучающей выборки, кроме него самого. Далее происходит следующее оценивание: верно ли, что по ближайшему окружению объекта x_i действительно можно правильно классифицировать объект x_i?

Несмотря на простоту реализации алгоритма и его широкий спектр применения в прикладных задачах, он имеет существенный недостаток при решении задачи в условиях дисбаланса отдельных признаков [5]. Такая проблематика встречается в разных прикладных задачах, однако в задаче прогнозирования точных зависимостей вероятность её присутствия крайне велика. Например, при прогнозировании медицинских заболеваний пациенты могут быть неравномерно распределены по половому признаку — и/или внутри категорий, разделенных по половому признаку, присутствует второй уровень дисбаланса по критерию, допустим, возраста пациента.

В таких ситуациях с целью повышения уровня обобщающей способности алгоритма каждый объект необходимо будет учитывать как соседа самому себе.

Вынужденность использования такого подхода влечет за собой смещение и оптимистичное занижение оценки числа ошибок (см. рисунок 1) [6].

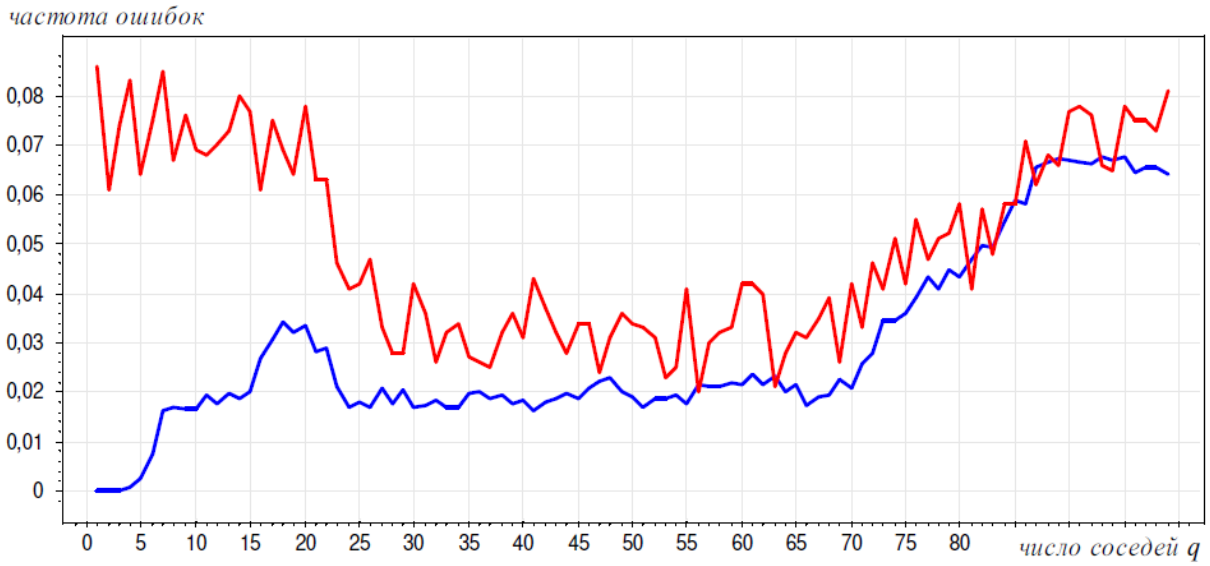

Рисунок 1. Зависимость смещённости числа ошибок в алгоритме k-ближайших соседей

На рисунке 1 по горизонтальной оси отложено число соседей в алгоритме k-ближайших соседей, а по вертикальной – частота ошибок. Синяя линия характеризует учет самого объекта в своей окрестности (среди ближайших объектов соседа есть он сам). При очень маленьком размере окрестности мы получаем практически нулевое значение ошибки (диапазон значений числа соседей от 0 до 5 на рисунке 1), в то время как «честная» классификация объекта (красная линия на рисунке 1) по его ближайшему соседу – это порядок ошибки в 8%. Получается, что ошибка функционирования алгоритма искусственного интеллекта при решении прикладной задачи искусственно занижается.

Таким образом, методика подсчета точности модели искусственного интеллекта, основанная на методе k-ближайших соседей, некорректно работает на наборе данных, признаки в котором не сбалансированы.

Список литературы:

1. Пылов, П. А. Изучение искусственного интеллекта на основе принципа интенсификации обучения / П. А. Пылов, Р. В. Майтак, А. В. Дягилева. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2024. – 172 с. – ISBN 978-5-9729-1594-1. – EDN YFPIKU.
2. Методы восстановления непараметрической регрессии в условиях несбалансированных данных / П. А. Пылов, Р. В. Майтак, А. В. Дягилева, А. Д. Салычева. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2024. – 192 с. – ISBN 978-5-9729-1856-0. – EDN AAJATW.
3. Протодьяконов, А. В. Асимптотический анализ поведения прикладных моделей машинного обучения : Учебное пособие / А. В. Протодьяконов, А. В. Дяги-

лева, П. А. Пылов. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 144 с. – ISBN 978-5-9729-1455-5. – EDN APHQME.
4. Дягилева, А. В. Корреляционный анализ успеваемости в дисциплинах базовой образовательной программы высшего образования и дополнительного образования: перспективы оптимизации учебного процесса / А. В. Дягилева, П. А. Пылов // Инновационная деятельность педагога: традиции и современность : Сборник материалов II Всероссийской научно-практической конференции, посвященной Году педагога и наставника, Владикавказ, 19 мая 2023 года. – Владикавказ: Северо-Осетинский государственный университет имени К.Л. Хетагурова, 2023. – С. 239-243. – EDN OZLILR.
5. Майтак, Р. В. Параметризация гиперпараметров в прикладных задачах машинного обучения на основе ядерных функций / Р. В. Майтак, П. А. Пылов // Россия молодая : СБОРНИК МАТЕРИАЛОВ XIV ВСЕРОССИЙСКОЙ, НАУЧНО-ПРАКТИЧЕСКОЙ КОНФЕРЕНЦИИ МОЛОДЫХ УЧЕНЫХ С МЕЖДУНАРОДНЫМ УЧАСТИЕМ, Кемерово, 18–21 апреля 2023 года. – Кемерово: Кузбасский государственный технический университет имени Т.Ф. Горбачева, 2023. – С. 31736.1-31736.6. – EDN QGJUMS.
6. Майтак, Р. В. Применение имитационной модели искусственного интеллекта для автоматизированной оптимизации процесса разработки угольных месторождений / Р. В. Майтак, П. А. Пылов, А. В. Дягилева // Математика и ИТ - вместе в цифровое будущее : сборник трудов II Молодежной школы, Нижний Новгород, 24–28 апреля 2023 года / Министерство науки и высшего образования РФ; Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского. – Нижний Новгород: Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского, 2023. – С. 52-56. – EDN IFHQUG.