

УДК 004.89

ПРИМЕНЕНИЕ МОДЕЛИ СЛУЧАЙНОГО ЛЕСА ДЛЯ ДЕТЕКЦИИ ВЫБРОСОВ В БОЛЬШИХ ДАННЫХ

Щукин Г.Р., студент гр. М3305, II курс
Университет ИТМО, г. Москва

Детекция выбросов в больших данных является важной задачей в различных областях, таких как финансы, медицина, маркетинг и кибербезопасность. Выбросы, или аномалии, могут указывать на ошибки в данных, неожиданные события или даже мошеннические действия.

Одним из эффективных методов для решения этой задачи является использование модели случайного леса. Случайный лес, благодаря своей структуре и свойствам, позволяет эффективно и надежно выявлять выбросы в больших объемах данных [1].

Что такое случайный лес? Случайный лес – это ансамблевый метод машинного обучения, который состоит из множества решающих деревьев. Каждый элемент случайного леса (решающее дерево) обучается на случайной выборке из исходного набора данных и использует случайную подвыборку признаков для разделения данных.

Результат каждого дерева комбинируется, чтобы улучшить общую производительность модели и снизить вероятность переобучения.

Применение случайного леса для детекции выбросов. Для выявления выбросов случайный лес можно использовать несколькими способами.

Один из наиболее распространенных методов – это использование алгоритма Isolation Forest. Isolation Forest специально разработан для детекции выбросов и основывается на принципе изоляции наблюдений [2].

Алгоритм Isolation Forest:

1. Создание множества деревьев.

Случайный лес создается из большого количества решающих деревьев, каждое из которых строится с использованием случайной подвыборки данных и признаков.

2. Изоляция наблюдений.

Каждое дерево в лесу изолирует наблюдения, разделяя их на основе случайно выбранных признаков. Выбросы, как правило, легче изолировать, так как они находятся далеко от большинства наблюдений и требуют меньше разделений для их изоляции.

3. Оценка аномальности.

Для каждого наблюдения вычисляется оценка аномальности на основе количества разделений, необходимых для его изоляции. Наблюдения с высокой оценкой аномальности классифицируются как выбросы [3].

Преимущества использования случайного леса для детекции выбросов:

1. Эффективность на больших данных.

Случайный лес хорошо масштабируется на большие объемы данных и эффективно работает в условиях высокой размерности.

2. Робастность к шуму.

За счет усреднения результатов множества деревьев случайный лес устойчив к шуму и выбросам в данных, что улучшает надежность детекции.

3. Отсутствие предположений о распределении данных.

Случайный лес не требует предположений о распределении данных, что делает его универсальным инструментом для различных типов данных [4].

4. Интерпретируемость.

Хотя случайный лес является ансамблевым методом, результаты могут быть интерпретированы, что помогает понять, какие признаки влияют на выявление выбросов.

Модель случайного леса, особенно в виде алгоритма Isolation Forest, представляет собой мощный и эффективный инструмент для детекции выбросов в больших данных.

Ее способность работать с большими объемами данных, устойчивость к шуму и гибкость в применении делают ее незаменимым инструментом в современном анализе данных.

Использование случайного леса для выявления аномалий продолжает развиваться, предоставляя новые возможности для обеспечения безопасности, повышения качества данных и принятия более информированных решений [5].

Список литературы:

1. Breiman, L. Random forests / L. Breiman // Machine Learning. – 2001. – Vol. 45. – P. 5–32. – DOI 10.1023/A:1010933404324.
2. Liaw, A. Classification and regression by randomForest / A. Liaw, M. Wiener // R News. – 2002. – Vol. 2, № 3. – P. 18–22.
3. Cutler, D. R. Random forests for classification in ecology / D. R. Cutler [et al.] // Ecology. – 2007. – Vol. 88, № 11. – P. 2783–2792. – DOI 10.1890/07-0539.1.
4. Fernández-Delgado, M. Do we need hundreds of classifiers to solve real world classification problems? / M. Fernández-Delgado [et al.] // Journal of Machine Learning Research. – 2014. – Vol. 15. – P. 3133–3181.

5. Oshiro, T. M. How many trees in a random forest? / T. M. Oshiro, P. S. Perez, J. A. Baranauskas // Machine Learning and Data Mining in Pattern Recognition. – Berlin : Springer, 2012. – P. 154–168. – DOI 10.1007/978-3-642-31537-4_13.