Гафуров И.А., Ситников С.Ю. Распознавание аномалий данных для прогнозирования отказов оборудования // Международный журнал информационных технологий и энергоэффективности. – 2023. – Т. 5 № 2(28) с. 9–12

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.89

РАСПОЗНАВАНИЕ АНОМАЛИЙ ДАННЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ ОТКАЗОВ ОБОРУДОВАНИЯ

[1]Гафуров И.А., [2]Ситников С.Ю.

Казанский государственный энергетический университет, Казань, Россия (420066, г. Казань, ул. Красносельская, 51, корп. Д);e-mail: [1] gafurov0ivan@gmail.com, [2] ssitnikov@mail.ru

Оборудование ТЭК генерирует большое количество данных, описывающих его состояние и показатели. Используя эти данные, можно выявить нормальные и аномальные показатели оборудования. Идея изучения нормального поведения с помощью реплицируемой нейронной сети и Isolation Forest применяется для выявления аномалий и прогноза отказов оборудования. Сравнивается, насколько хороши методы обнаружения аномалий и как это применимо к данным, собранным на оборудовании ТЭК. Можно сделать вывод, что Isolation Forest превосходит реплицированную нейронную сеть в данном виде анализа.

Ключевые слова: большие данные, отказоустойчивость, масштабируемость, обнаружение аномалий.

RECOGNITION OF DATA ANOMALIES FOR PREDICTION OF EQUIPMENT FAILURES

[1] Gafurov I.A., [2] Sitnikov S.Y.

Kazan State Power Engineering University, Kazan, Russia (420066, Kazan, Krasnoselskaya str., 51, bldg. D), e-mail: [1] gafurov0ivan@gmail.com, [2]ssitnikov@mail.ru

Fuel and energy complex equipment generates a large amount of data describing its condition and performance. Using this data, you can identify normal and abnormal equipment performance. The idea of studying normal behavior using a replicated neural network and Isolation Forest is used to detect anomalies and predict equipment failures. It compares how good anomaly detection methods are and how it applies to data collected on fuel and energy equipment. It can be concluded that the Isolation Forest is superior to the replicated neural network in this type of analysis.

Keywords: big data, fault tolerance, scalability, anomaly detection.

Выявление аномалий применяются в различных областях например, чтобы распознать атаки на сеть (обнаружение вторжений), обнаружение финансовых мошенничеств и в медицине для выявления болезней. Выявление аномалий в показаниях датчика оборудования позволяет спрогнозировать неисправность до того, как оборудование выйдет из строя.

Основанные на моделировании методы обнаружения аномалий это реплекативные нейронные сети (ReplNN). Более новым методом является изолированный лес (Isolation Forest).

В реплекативных нейронных сетях выделяет сферу всех значений датчиков, которая включает все нормальные данные, найденные во время обучения. При анализе показания датчиков за пределами этой сферы определяются как аномальные.

Во время обучения репликативной нейронной сети функция предсказания $f(x)$ определяется таким образом, что разница между точкой обучения x ∈ R и ее результатами $f(x) = \tilde{x}, \tilde{x} \in R^d$ сведится к минимуму для всех x ∈ T , где T описывает обучающие данные, а d количество итераций обучения. Веса настроены с нормальными данными, и в при анализе ошибка при восстановлении точки тестовых данных xtest характеризует ее аномалию. Ошибка восстановления рассчитывается как среднеквадратическая разница между исходным значением и его восстановление: $||x_{test} - \tilde{x}_{test}||^2_2$

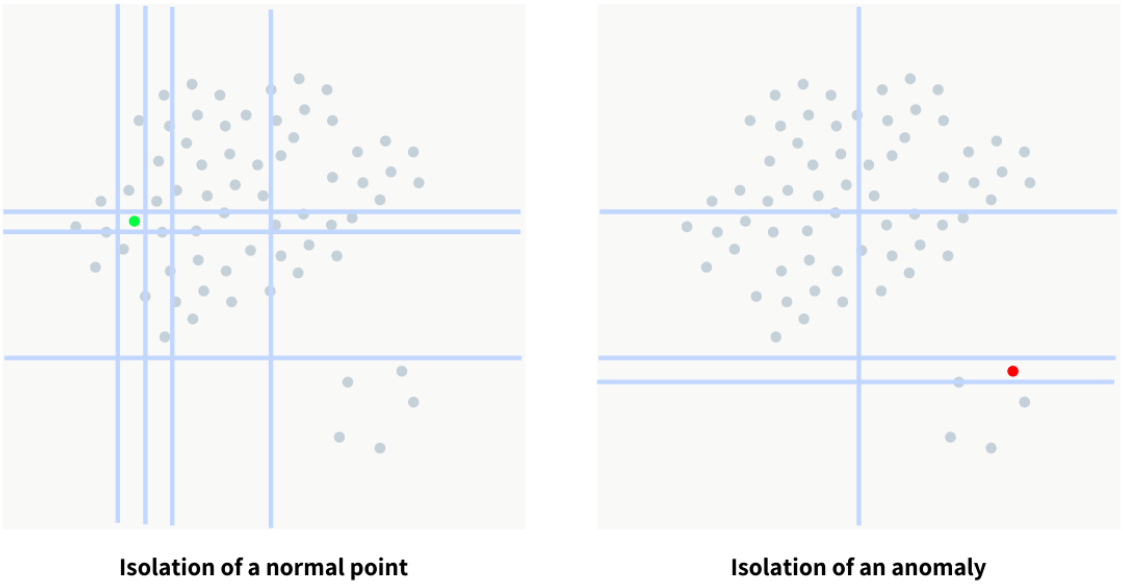

Рисунок 1 – Изоляция нормальной точки и аномалии

Подход Isolation Forest предполагает, что аномалии легче изолировать от остальных данных, чем обычные экземпляры (Рисунок 1).

Многие бинарные деревья генерируются с использованием выборки обучающих данных T. Таким образом, функция и значение разделения выбираются случайным образом (равномерное распределение). Единственными параметрами, которые необходимо определить, являются количество деревьев (t) и размер выборки (ψ).

Дерево, которое наиболее эффективно выявляет нормальные и ненормальные данные содержит ожидаемую длину пути аномальной точки данных короче, чем длина пути нормальных точек. Соответственно для примера:

$$P(T_1) < P(T_2) = \Rightarrow E(h_2) < E(h_0) < E(h_1),$$

где $P(T_i)$ описывает вероятность дерева T_i и $E(h_i)$ как ожидаемая длина пути (h_i) точки данных x_i, $i \in \{0,1,2\}$:

$$E(h_i) = P(h(x_i) = 1) * 1 + P(h(x_i) = 2) * 2.$$

Средняя длина пути точки тестовых данных.

Обычные сгенерированные деревья *xtest* используются для оценки аномалий. Чем меньше значение, тем выше вероятность быть аномалией.

Для обнаружения аномалий в данных датчика оборудования, используется концепция одноклассовой классификации. Поэтому элементы данных собираются на серверной части, и предполагая, что большинство из них являются нормальными, изучается эталонная модель. На этапе анализа модель применяется в системе анализа больших данных для выявления отклонения от нормального поведения.

Мы оценили указанные методы обнаружения аномалий в неструктурированных данных датчиков оборудования, используя разные наборы данных.

На основе данных, собранных на объекте ТЭК, обучили модель нейронной сети для фиксации представления нормальных данных. Также ровели обучение использую данные с аномалиями, такие как авария, полная поломка или неисправность. Этимассивы данных были полностью промаркированы.

Анализируемые данные — это сигналы датчиков, передаваемые через внутреннюю систему оборудования и сервера SCADA, более конкретно через сеть контроллеров Siemmens. Эти сигналы, такие как скорость, продольное/поперечное ускорение, давление, температура описывает операции оборудования [1-2].

Наша оценка показывает, что оценка аномалий с помощью алгоритма Isolation Forest может быть использована для идентификации очень нестандартных ситуаций и дает лучшие результаты, чем реплицируемая нейронная сеть. Определение основанной на весах границы, различающей нормальное и ненормальное поведение, сложно и зависит от варианта использования. Неправильно классифицированные данные затем могут быть нивелированы при комплексном анализе всех показаний датчиков, собранных в течение одной операции оборудования, а также объединении данных операций во множества.

Следующими шагами будут дальнейшие улучшения модели. Поскольку оборудование ТЭК производит зависящее от времени данные, необходим пересмотр весов для настройки указанного алгоритма. для учета характеристик, описывающих временной ряд [3].

Данную модель можно успешно использовать для прогноза возможности отказов, простоев и неполадок, что позволит планировать мероприятия по обслуживанию оборудования для предотвращения простоев, уменьшить затраты на его обслуживание и продлить срок службы.

Список литературы

1. Bart Baesens, Analytics in a Big Data World: The Essential Guide to Data Science and its Applications. New Jersey: John Wiley & Sons, Inc., 2014. — pp. 35-71.
2. Marz N., Warren J. Big Data: Principles and Best Practices of Scalable Realtime. New York: Manning Publications Co., 2015. — pp. 225-241.
3. Mandic D.P., Chambers J.A. Recurrent neural networks for prediction. NewYork: John Wiley and Sons, Inc., 2001. pp. 171-198.

References

1. Bart Baesens, Analytics in a Big Data World: The Essential Guide to Data Science and its Applications. New Jersey: John Wiley & Sons, Inc., 2014. — pp. 35-71.

2. Marz N., Warren J. Big Data: Principles and Best Practices of Scalable Realtime. New York: Manning Publications Co., 2015. — pp. 225-241.
3. Mandic D.P., Chambers J.A. Recurrent neural networks for prediction. NewYork: John Wiley and Sons, Inc., 2001. pp. 171-198.