

УДК 621.65:004.85

МЕТОДИКА ВЫБОРА ГИПЕРПАРАМЕТРОВ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ОТКАЗОВ НАСОСНОГО ОБОРУДОВАНИЯ ПО ДАНЫМ ВИБРОДИАГНОСТИКИ

Севостьянова Л.И.

студент кафедры автоматизации технологических процессов и производств СамГТУ

e-mail: lsevestyanova@mail.ru

Аннотация

Рассматривается методика выбора гиперпараметров моделей машинного обучения для диагностики состояния и оценки риска отказов насосных агрегатов по вибрационным сигналам. Предложены этапы регистрации данных, сегментации, получения временных и спектральных признаков, формирования целевой переменной и проверки качества без утечки информации между измерениями. Для сопоставления выбраны метод опорных векторов, случайный лес и градиентный бустинг XGBoost. Описаны пространства поиска гиперпараметров, процедура вложенной групповой проверки, показатели качества для редких отказов и правила выбора эксплуатационного порога. Разграничены задачи распознавания уже наблюдаемого дефекта и прогнозирования будущего отказа. Работа представляет методическое исследование; результаты натурных испытаний, которые не проводились, не приводятся.

Ключевые слова: *насосное оборудование, вибродиагностика, прогнозирование отказов, машинное обучение, гиперпараметры, метод опорных векторов, случайный лес, XGBoost, утечка данных.*

HYPERPARAMETER SELECTION METHODOLOGY FOR MACHINE LEARNING MODELS PREDICTING PUMP EQUIPMENT FAILURES FROM VIBRATION DATA

Sevestyanova L.I.

Student of the Department of Automation of Technological Processes and Productions, Samara State Technical University

e-mail: lsevestyanova@mail.ru

Abstract

This paper proposes a reproducible hyperparameter selection methodology for machine learning models used to assess pump condition and predict failures from vibration signals. The workflow covers acquisition, windowing, feature extraction, labeling, leakage-resistant validation and operational threshold selection. Support vector machines, random forests and XGBoost are examined under a common search budget. Particular attention is given to group-aware and chronological splits, rare failure events and the distinction between fault diagnosis and genuine advance failure prediction. The work is methodological and does not claim unperformed laboratory experiments.

Keywords: *centrifugal pump, vibration monitoring, predictive maintenance, machine learning, hyperparameter tuning, support vector machine, random forest, XGBoost.*

Введение

Отказы насосных агрегатов приводят к остановкам технологических линий, изменению давления и расхода, внеплановым ремонтам и дополнительным затратам. Особенно чувствительны к ухудшению технического состояния опорные подшипники, муфты, рабочие колеса и уплотнения. Вибрационный сигнал отражает изменения в работе вращающихся элементов, однако его форма определяется не только дефектом, но и частотой вращения, гидравлическим режимом, жесткостью фундамента, местом крепления датчика и уровнем фоновых помех. Поэтому достоверное использование машинного обучения требует сначала инженерного описания измерений, а уже затем настройки классификатора [1, 2].

На практике нередко сравнивают модели по доле правильно распознанных коротких фрагментов записи. Если фрагменты одной и той же виброграммы случайным образом попадают одновременно в обучающую и тестовую выборки, полученная точность может характеризовать сходство соседних окон, а не способность распознавать состояние другого агрегата либо будущую неисправность. Не менее существенна подмена прогноза отказа диагностикой заранее сформированного дефекта. Отличить эти постановки необходимо до выбора гиперпараметров.

Цель статьи – разработать методически обоснованный и воспроизводимый порядок выбора гиперпараметров трех моделей для работы с вибродиагностическими данными насосного оборудования. Предлагаемый порядок включает подготовку признаков, разделение наблюдений, поиск конфигурации, настройку порога тревоги и сравнение с простым диагностическим базисом. Новизна методического подхода состоит в совместном учете источника вибрационных окон, временного горизонта прогноза и эксплуатационной стоимости пропущенного отказа.

Методы и методики

Рассмотрим насосный агрегат с электродвигателем, соединительной муфтой и вращающейся частью насоса. Для него регистрируются ускорения вибрации в радиальном и, при наличии датчика, осевом направлениях. Запись должна сопровождаться идентификатором агрегата, временем наблюдения, частотой дискретизации f_s , частотой вращения n , нагрузкой, температурой, состоянием после ремонта и отметками технических событий. Без этих сопроводительных признаков одинаковые по амплитуде сигналы при различных режимах невозможно корректно интерпретировать.

Основной сигнал обозначим $a(t)$, где t – время. Измерение разбивают на окна продолжительностью T_v при шаге ΔT . Число отсчетов в окне равно $M = f_s T_v$; перекрытие окон следует документировать. Для воспроизводимого демонстрационного протокола допустимо начать с $T_v = 1$ с, перекрытия 50 % и частоты, соответствующей характеристикам реально установленного датчика. Эти значения являются стартовыми настройками исследования, а не универсальными нормативами для всех насосов.

$$M = f_s T_v, \quad x_k = [a(k\Delta T), \dots, a(k\Delta T + T_v)], \quad (1)$$

При изучении состояния насоса необходимо разделять гидравлические и механические причины повышенной вибрации. Кавитация способна изменять широкополосную составляющую сигнала, дисбаланс – составляющую на частоте вращения, а повреждение подшипника – импульсность и характерные спектральные линии. Эти связи следует рассматривать как диагностические гипотезы: отдельный признак редко является достаточным доказательством конкретного дефекта. Полезно фиксировать режим расхода и давления, если он доступен, либо хотя бы категорию эксплуатационного режима.

В открытом доступе существуют реальные вибрационные данные подшипников. Case Western Reserve University предоставляет записи исправных и поврежденных подшипников испытательного электродвигателя, в том числе при частотах дискретизации 12 и 48 кГц [3]. Набор Paderborn University включает вибрацию, ток и сопутствующие условия для исправных и поврежденных подшипников при четырех режимах работы [4]. Такие наборы пригодны для отладки вычислительного конвейера, но они не являются прямыми измерениями отказов насосов: перенос обученных на них моделей требует отдельного подтверждения.

Прежде чем рассчитывать признаки, из данных исключают записи с известным отказом датчика, нарушением синхронизации, насыщением измерительного тракта и неустановленным режимом агрегата. Спорные интервалы не следует автоматически относить к исправному состоянию. Для каждой записи сохраняют исходный файл и журнал исключений, иначе эксперимент нельзя проверить независимо. Фильтрация и нормализация должны быть описаны с частотами среза и способом оценки коэффициентов.

Первичную матрицу признаков составляют статистики виброускорения: среднеквадратическое значение (RMS), пиковая амплитуда, размах, коэффициент эксцесса, коэффициент амплитуды (crest factor), а также энергия выбранных спектральных полос. RMS

отражает общий уровень колебаний, тогда как эксцесс и crest factor помогают выделять короткие ударные события. Спектральные признаки рассчитывают на основе дискретного преобразования Фурье с одинаковым окном и масштабированием для всех наблюдений [1, 2].

$$RMS(x) = \sqrt{\left[\left(\frac{1}{M}\right) \sum_{i=1}^M x_i^2\right]}, \quad CF(x) = \frac{\max_i |x_i|}{RMS(x)}, \quad (2)$$

Для вращающегося оборудования целесообразно нормировать частотные компоненты на частоту вращения или применять порядок гармоник (order analysis), если обороты заметно меняются. При фиксированных оборотах можно сравнивать спектральную энергию вокруг первой и второй гармоник частоты вращения, полосы высокой частоты и огибающую вибросигнала. Однако частоты дефектов подшипников зависят от геометрии узла: без сведений о типоразмере подшипника нельзя приписывать каждой спектральной линии определенный механический смысл.

Целевую переменную задают в зависимости от исследовательского вопроса. Для диагностики $y = 1$ означает документированное наличие дефекта во время регистрации окна. Для прогнозирования отказа требуется журнал событий с моментом отказа t^f и прогнозный горизонт H . В этом случае наблюдение в момент t маркируют положительным, когда в пределах $(t, t + H]$ действительно произошел заданный отказ; вводится также минимальный интервал упреждения h , чтобы алгоритм не получал преимущество от измерений непосредственно в момент аварии. В отсутствие истории реальных отказов возможна только постановка диагностики или исследования аномалий, но не проверенный прогноз момента отказа.

$$y(t; H, h) = 1\{t + h < t^f \leq t + H\}, \quad (3)$$

Записи после технического обслуживания следует рассматривать как новую эксплуатационную фазу, а незавершенные наблюдения без последующего события – как цензурированные, если полный горизонт H не наблюдался. Запрещается превращать такие наблюдения в отрицательные примеры лишь потому, что в доступном файле не указан отказ. Разметку выполняют до оконного преобразования, но при разделении данных учитывают происхождение всех окон из одной непрерывной записи.

Для инженерной выборки умеренного объема рационально использовать несколько разнородных алгоритмов. В качестве первого метода рассматривается машина опорных векторов с радиальной базисной функцией (SVM-RBF). Она строит нелинейную разделяющую поверхность и может эффективно работать на стандартизованных статистических признаках. Главные гиперпараметры – штраф C за ошибки классификации и коэффициент γ , определяющий ширину ядра. При слишком большом C модель способна переобучиться, а слишком большой γ приводит к узким областям влияния обучающих наблюдений [5].

Второй вариант – случайный лес (Random Forest), то есть ансамбль деревьев решений, обучаемых на случайных подвыборках объектов и признаков [6]. Метод способен учитывать нелинейные взаимодействия и обычно не требует масштабирования признаков. Основные настраиваемые величины: число деревьев, максимальная глубина, минимальный размер листа, число рассматриваемых признаков при разбиении и вес редкого класса. Для данных небольшой размерности избыточно глубокие деревья могут запоминать особенности конкретной измерительной серии.

Третий вариант – градиентный бустинг деревьев XGBoost, последовательно исправляющий ошибки предшествующих моделей с регуляризацией сложности [7]. Существенны скорость обучения, количество деревьев, их глубина, минимальная масса листа, доли подвыборок строк и признаков, а также коэффициенты регуляризации. При недостаточном числе независимых экспериментальных серий такой алгоритм способен давать очень высокую внутреннюю оценку, которая не подтверждается на другом насосе.

Рекомендуемые исходные пространства поиска гиперпараметров систематизированы в табл. 1.

Таблица 1 – Рекомендуемые исходные пространства поиска гиперпараметров

Модель	Гиперпараметры	Начальный диапазон поиска
SVM-RBF	C; γ ; вес класса	C: $10^{-2} \dots 10^3$; γ : $10^{-4} \dots 10^1$; balanced/без весов
Random Forest	Число деревьев; глубина; min samples leaf; max features	200-800; 4-24/без ограничения; 1-10; sqrt/0,5/1,0
XGBoost	n_estimators; max_depth; learning_rate; subsample; colsample_bytree	150-800; 2-8; 0,01-0,2; 0,6-1,0; 0,6-1,0

Диапазоны таблицы являются стартовым проектом поиска, а не эмпирически найденными оптимальными настройками. Их расширяют только при наличии достаточного числа независимых наблюдений и признаков недообучения или переобучения. Для SVM параметры C и γ следует выбирать в логарифмической шкале. Для бустинга число итераций разумно согласовать с ранней остановкой, основанной исключительно на внутренней проверочной подвыборке.

Результаты

На первом этапе данные разбивают на обучающую и окончательную тестовую части на уровне независимых единиц наблюдения. Такой единицей может быть отдельный насос, испытательный прогон, ремонтный цикл либо непрерывная запись, но не одномерное окно из нее. При оценке переноса между машинами желательно полностью исключить выбранные агрегаты из обучения. При оценке будущих событий на том же агрегате применяют хронологическое разделение с буфером между периодами, перекрывающим длительность окна и при необходимости горизонт маркировки.

На втором этапе внутри обучающей части проводят вложенную проверку. Внешние разбиения служат для честной оценки процедуры настройки; во внутренних разбиениях выполняют поиск гиперпараметров, предварительную стандартизацию и, если допустимо, отбор признаков. Все статистики стандартизатора вычисляются только по тренировочной части каждого внутреннего разбиения. Перенос нормализации, рассчитанной по полному массиву, в начало конвейера является типичной утечкой информации [8].

На третьем этапе всем сравниваемым моделям выделяют сопоставимый вычислительный бюджет: например, по 40 случайных конфигураций при одинаковом числе внутренних разбиений. Случайный поиск полезен как прозрачный ориентир: показано, что он во многих задачах рациональнее полного перебора сетки при том же бюджете [9]. Границы поиска и генератор случайных чисел фиксируют до просмотра итоговых результатов, а не корректируют после оценки тестовой выборки.

Для каждой конфигурации θ вычисляют среднее значение целевой метрики по K внутренним разбиениям. Выбор основывают не только на наилучшем среднем, но и на разбросе качества: при почти равных оценках предпочтительна более простая модель с меньшим временем предсказания. При больших различиях между рабочими режимами полезно дополнительно рассчитывать минимальную оценку качества по отдельным режимам, чтобы общая метрика не скрывала провал при редкой нагрузке.

$$\theta^* = \arg \max_{\theta \in \Theta} \left(\frac{1}{K} \sum_{k=1}^K Q_k(\theta) \right), \quad (4)$$

где Θ – заранее объявленное пространство поиска;

Q_k – значение целевой метрики на k-м внутреннем валидационном блоке.

При сильном дисбалансе классов главной метрикой можно назначить площадь под кривой precision-recall (PR-AUC, average precision), дополнив ее полнотой при заданном ограничении на число ложных тревог. Точность ассурасу не должна быть единственным критерием: модель, почти всегда предсказывающая исправность, может демонстрировать высокий процент правильных ответов и пропускать опасные события.

Если внешние разбиения проводятся на разных насосах, результат отражает устойчивость выбора к изменению агрегата. Если они сформированы по временным блокам

одного насоса, вывод относится к прогнозу на более поздний период эксплуатации. Смешивать эти оценки в одном числе нельзя. После завершения выбора конфигурации окончательную модель обучают на всей доступной обучающей части и один раз испытывают на ранее закрытом тестовом блоке. Тест не используется для пересмотра диапазонов или выбора порога.

Для оценки распознавания отказов определяют TP – правильно выявленные события, FN – пропущенные, FP – ложные тревоги, TN – верно определенное исправное состояние. Из них рассчитывают $precision = TP/(TP+FP)$, $recall = TP/(TP+FN)$, F1 и PR-AUC. Важно заранее решить, относится ли единица учета к окну, событию или агрегату-часу. Один длительный отказ, давший тысячу положительных окон, не эквивалентен тысяче независимых аварий.

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}, \quad (5)$$

Для реального обслуживания оборудования дополнительно необходимы число ложных тревог на 1000 часов эксплуатации, доля выявленных отдельных отказов и распределение времени предупреждения до остановки. Порог вероятности тревоги выбирают на внутренних валидационных данных с учетом цены пропуска отказа и расходов на проверку. Не следует назначать порог 0,5 автоматически: оптимальная эксплуатационная точка зависит от распространенности дефектов и доступности ремонтного персонала.

Если выход модели трактуется как вероятность, проверяют ее калибровку, например, кривой надежности и Brier score. Калибровку выполняют отдельно от обучения классификатора и не на финальном тесте. Для временных потоков целесообразно требовать несколько последовательных превышений порога перед подачей сигнала тревоги; это снижает реакцию на единичные выбросы, но способно уменьшить время упреждения. Такой фильтр рассматривается как часть полной диагностической системы (табл. 2).

Таблица 2 – Протокол сопоставления моделей на едином наборе измерений

Элемент проверки	Единое требование	Что отражает
Разделение	По насосу/прогону или по времени с буфером	Способность к переносу, отсутствие соседних окон в тесте
Настройка	Одинаковый бюджет поиска и внутренние блоки	Корректность сравнения моделей
Основная метрика	PR-AUC; дополнительно recall	Работа при редких отказах
Эксплуатационная метрика	Ложные тревоги/1000 ч; упреждение	Полезность для техобслуживания
Отчетность	Интервалы неопределенности; время обучения и инференса	Стабильность и вычислительные затраты

Таблица показывает, какие характеристики необходимо получить из реальных испытаний; она не содержит вымышленных численных результатов. Для анализа неопределенности предпочтительно повторно выбирать независимые агрегаты или испытательные серии, а не отдельные коррелированные окна. При малом числе реальных отказов интервал неопределенности может оказаться широким даже при высоком значении PR-AUC, и это ограничение следует явно указывать.

Обсуждение

Практический эксперимент удобно строить в два этапа. На лабораторном стенде насосной установки регистрируют серии при исправном состоянии и нескольких устойчивых рабочих режимах; далее собирают записи при подтвержденных дефектах либо используют документированные записи предаварийной эксплуатации. Искусственное создание опасных отказов на рабочем насосе недопустимо без специально оборудованного и безопасного испытательного стенда. Для каждого опыта фиксируют настройки регистрации и заключение специалиста о техническом состоянии узла.

При недостатке локальных данных следует разделить учебную проверку вычислительного конвейера на открытых подшипниковых массивах [3, 4] и собственно

насосное исследование. Классификатор, хорошо распознающий искусственно поврежденные подшипники электродвигателя, не становится автоматически валидированным прогнозом кавитации, повреждений крыльчатки или потери герметичности насоса. Даже при общей механической природе части неисправностей отличаются частотные характеристики, места установки датчиков и классы событий.

Аналитически SVM является разумным ориентиром при компактном векторе тщательно отобранных признаков; его чувствительность к масштабу признаков требует включать StandardScaler внутри проверочного конвейера. Random Forest удобен как устойчивый базовый ансамбль с относительно небольшим объемом предварительных преобразований и возможностью оценить вклад признаков, хотя такая оценка сама по себе не доказывает причинность. XGBoost позволяет учитывать сложные взаимодействия, но требует особенно аккуратной регулировки глубины и темпа обучения, чтобы не подстраиваться под конкретный прогон [5-7].

Сравнение моделей следует выполнять по результатам одного и того же разбиения и одинаковым критериям отбора. Нельзя сделать вывод «XGBoost точнее SVM» только из литературных результатов на разных наборах: изменяются частота дискретизации, баланс классов, набор признаков и даже смысл слова «отказ». Конечная рекомендация должна исходить из собственных внешних проверочных блоков, а при равенстве качества учитывать скорость расчета на промышленном контроллере, сложность сопровождения и возможность обнаружить деградацию входных сигналов.

Отдельное направление – контроль изменения режима данных после внедрения. Со временем меняются упругость опор, балансировка после ремонта, датчики и характеристики продукта, перемещаемого насосом. Следовательно, алгоритм должен сопровождаться проверкой статистического сдвига распределения признаков и журналом ложных тревог. Перенастройку гиперпараметров проводят на накопленных, заново размеченных данных, сохраняя неизменный независимый контрольный период для последующей проверки.

Важнейшее методическое ограничение касается прогнозного горизонта. Чем длиннее H , тем полезнее предупреждение ремонтной службе, но тем неопределеннее связь текущей вибрации с будущим отказом. В статье предлагается сравнивать по меньшей мере два заранее зафиксированных горизонта, когда доступны журналы отказов, и публиковать качество по каждому отдельно. Нельзя выбирать наиболее удачный горизонт после просмотра тестовой выборки: это еще одна форма неявной настройки по тесту.

Заключение

Предложена воспроизводимая методика выбора гиперпараметров SVM-RBF, Random Forest и XGBoost для обработки вибрационных сигналов насосных агрегатов. Она охватывает документирование режима измерений, выделение статистических и спектральных признаков, строгую разметку событий, групповое либо хронологическое разделение данных и вложенную процедуру поиска параметров. В отличие от случайного разбиения коротких фрагментов предложенная схема ориентирована на проверку способности модели работать на независимых сериях и будущих интервалах.

Определены стартовые пространства поиска, метрики качества для редких отказов и порядок настройки эксплуатационного порога предупреждения. Особое значение имеет различие между диагностикой уже существующего дефекта и прогнозированием наступления будущего отказа.

Литература

1. Randall R. B. Vibration-based Condition Monitoring: Industrial, Aerospace and Automotive Applications. Chichester: John Wiley & Sons, 2011. 309 p.
2. Jardine A. K. S., Lin D., Banjevic D. A review on machinery diagnostics and prognostics implementing condition-based maintenance // Mechanical Systems and Signal Processing. 2006. Vol. 20, № 7. PP. 1483-1510.
3. Case Western Reserve University. Bearing Data Center: Download a Data File. URL: <https://engineering.case.edu/bearingdatacenter/download-data-file> (дата обращения: 21.10.2026).

4. Paderborn University. Bearing DataCenter: Data Sets and Download. URL: <https://mb.uni-paderborn.de/en/kat/research/bearing-datacenter/data-sets-and-download> (дата обращения: 21.10.2026).
5. Cortes C., Vapnik V. Support-vector networks // Machine Learning. 1995. Vol. 20. PP. 273-297.
6. Breiman L. Random forests // Machine Learning. 2001. Vol. 45. PP. 5-32.
7. Chen T., Guestrin C. XGBoost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016. PP. 785-794.
8. Kaufman S., Rosset S., Perlich C., Stitelman O. Leakage in data mining: formulation, detection, and avoidance // ACM Transactions on Knowledge Discovery from Data. 2012. Vol. 6, № 4. Art. 15.
9. Bergstra J., Bengio Y. Random search for hyper-parameter optimization // Journal of Machine Learning Research. 2012. Vol. 13. PP. 281-305.