УДК 004.056 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait.2020.2/2921
Поступила в редакцию 03.05.2020
Подписана в печать 15.06.2020

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПРОЦЕДУР КЛАСТЕРИЗАЦИИ ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ ПОКАЗАТЕЛЕЙ, ХАРАКТЕРИЗУЮЩИХ ФУНКЦИОНИРОВАНИЕ КОНТАКТНОЙ СЕТИ ЖЕЛЕЗНЫХ ДОРОГ

2020 Т. А. Моисеева✉, Т. М. Леденева

Воронежский государственный университет
Университетская пл., 1, 394018 Воронеж, Российская Федерация

Аннотация. В данной статье рассматривается актуальная задача обнаружения аномалий в работе оборудования и её решение методами кластеризации на примере анализа работы оборудования в железнодорожной области. Приводится описание различных стратегий обслуживания и выделяется проактивная стратегия обслуживания в качестве наиболее перспективной. Рассматриваются основные компоненты и задачи системы проактивной стратегии обслуживания в применении к железнодорожному сектору. Подробно рассматривается модуль диагностики и ставится задача обнаружения аномалий на примере значений выборки электрических параметров, характеризующих функционирование контактной сети. Предполагается, что исходными данными для задачи обнаружения аномалий являются временные ряды значений электрических сигналов. Для предварительной обработки данных выбраны методы спектрального анализа: оценка спектральной плотности мощности, и используется метрика, основанная на периодограмме. Для обработки данных используется метод временного окна. Производится сравнение работы метода опорных векторов и метода кластеризации К-средних на тестовых данных и оценивается доля правильных ответов. Были подобраны оптимальные параметры процедур.
Ключевые слова: обнаружение аномалий, обработка временных рядов, кластеризация, метод опорных векторов, проактивная стратегия обслуживания.

ВВЕДЕНИЕ

Техническое обслуживание железнодорожного сектора обладает рядом специфических особенностей и обусловливает выполнение таких требований, как: повышенные требования к безопасности, доступности и надежности, способность функционировать в условиях интенсивного трафика, увеличенных нагрузок и ограниченного времени на обслуживание [1]. Железнодорожная инфраструктура обладает протяженным характером и эксплуатируется в сложных, изменяющихся условиях, что способно вызвать усиленный износ оборудования и непредсказуемые отказы. С увеличением скоростного и грузового движения становится необходимым оптимизировать все процессы для сокращения времени и расходов на ремонт и обслуживание. В настоящее время усовершенствование систем обслуживания является частью программы инноваций в РЖД, а также является целью для многих европейских государств, таких, как Бельгия (Infrabel), Китай, Германия (Deutsche Bahn), Италия (Trenitalia) в планах модернизации в связи с усиленным развитием высокоскоростного движения в этих странах [2].

Правильно выбранная стратегия технического обслуживания подвижного состава и железнодорожной инфраструктуры имеет критическое значение в обеспечении безопасности, стабильности, надежности и эффективности пассажирских и грузоперевозок. В целом, стратегии технического обслу-

✉ Моисеева Татьяна Александровна
e-mail: lina.inverse1995@mail.ru

живания и ремонтов могут быть разделены на несколько групп [3]: стратегия ремонтов после отказа, стратегия планово-предупредительных ремонтов, стратегия ремонтов по состоянию, проактивная стратегия технического обслуживания и ремонтов, или диагностическое обслуживание (predictive maintenance, PdM). Выбор стратегии обслуживания, как правило, заключается в нахождении компромисса между максимальным сроком использования детали и риском простоя оборудования из-за отказа. На сегодняшний день наиболее приемлемой в железнодорожной области по оценкам специалистов является проактивная стратегия [4, 5]. Решения, принимаемые в рамках проактивной стратегии обслуживания, обоснованы текущим состоянием оборудования и прогнозом его состояния в будущем. Данные, поступающие с оборудованных сенсорами и датчиками деталей, подлежащих наблюдению, обрабатываются с целью определения оставшегося времени жизни детали, выявления и устранения отклонений и неисправностей в работе механизмов, и являются источником данных для системы поддержки принятия решений.

Внедрение проактивной стратегии обслуживания предполагает автоматизацию процессов, разработку новых высокотехнологичных инструментов для диагностики состояния оборудования и создание на этой основе систем поддержки принятия решений (СППР) для оптимизации технических работ.

Ядром СППР является диагностический модуль, который формирует исходную информацию для анализа и формирования возможных решений. Наиболее распространенным подходом к решению задачи диагностики является использование кластерных процедур для выявления аномалий в наблюдаемых данных. Анализ существующих исследований позволил выделить четыре группы подходов, применяющихся для решения данного класса задач: подходы, основанные на статистических тестах [6], модельный подход [7, 8], метрические методы [9] и методы машинного обучения [10, 11]. Кроме того, для решения задачи обнаружения аномалий существует возможность создания ансамбля алгоритмов, как правило, принадлежащих разным группам [12]. Теоретическое обоснование использования ансамбля алгоритмов представлено в [13]. Многими исследователями отмечается, что методы кластеризации с высокой точностью позволяют решать задачи обнаружения аномалий при функционировании контактной сети железных дорог [14, 15, 16, 17], в связи с чем была поставлена задача проведения сравнительного анализа процедур кластеризации для диагностики состояния контактной сети на основе доступных измерений электрических параметров тока.

Цель статьи заключается в выборе и сравнительном анализе процедур кластеризации для диагностики аномальных значений временных рядов показателей, характеризующих функционирование контактной сети железных дорог. Результаты анализа позволят выбрать алгоритм из группы алгоритмов кластеризации, который может использоваться в качестве основного для дальнейших исследований увеличения точности обнаружения аномалий в области диагностики контактной сети путем создания ансамбля алгоритмов из разных групп.

1. МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

1. 1. Постановка задачи обнаружения аномалий в функционировании оборудования

Исходная информация для решения задачи обнаружения аномалий задается в форме временных рядов. В качестве наблюдаемых параметров выступают мгновенные значения силы тока i (А), напряжения u (В) и освещенности (Вт/м^2).

Соответствующие временные ряды i и u порождаются следующими гармоническими функциями:

$$u = U_A \sin(\omega t + \psi) \qquad (1.1)$$

$$i = I_A \sin(\omega t + \psi) \qquad (1.2)$$

В результате воздействия нестационарных тяговых нагрузок электрическая система переменного тока характеризуется значи-

тельными колебаниями активной и реактивной мощностей, что вызывает провалы и выбросы питающего напряжения и возникновение гармоник [18, 19]. Графики значений *u* и *i* в тяговой сети переменного электрического тока представлены на рис. 1.

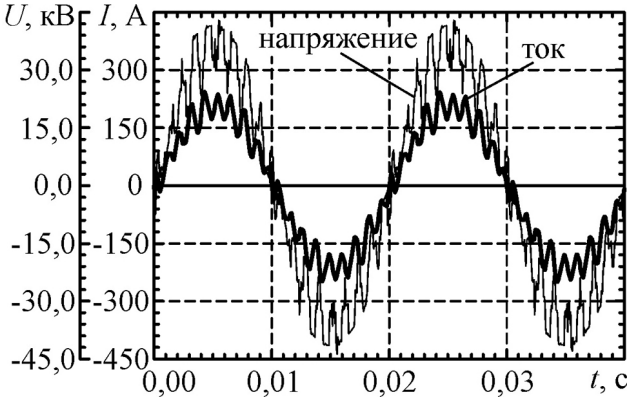

Рис. 1. Примеры графиков значений u и i
[Fig. 1. Examples of voltage and current values graphs]

В контексте поставленной задачи последовательность значений освещенности является источником целевых значений, позволяя использовать алгоритмы обучения с учителем.

Предположим, что отказ оборудования обусловлен появлением аномальных значений в некотором временном ряду. Задача заключается в определении соответствующего показателя.

Каждому моменту времени поставим в соответствие некоторое число $a_i \in [0,1]$, называемое рейтингом аномальности, показывающее, насколько нетипичны в данный момент значения временного ряда. В простейшем случае рейтинг аномальности принимает значения из $\{0,1\}$, т. е. является индикатором аномальности конкретного значения временного ряда.

1. 2. Общая схема решения

1.2.1. Обработка входных данных

Для обнаружения аномалий в исходном временном ряду воспользуемся методом временного окна. Оценка аномальности присваивается каждому окну временного ряда. Размер окна обозначим параметром *m*. Данный параметр требует тщательной настройки. Ввиду характера входных данных для сокращения размерности можно использовать методы спектрального анализа для выделения значимой информации. Воспользуется оценкой спектральной плотности мощности (СПМ) [20] для каждого временного окна. Обработка входных данных состоит из следующих шагов:

Сегментация исходного временного ряда на *n* непересекающихся временных рядов с размером окна *m*:

$$\left\{x_i\left(k\Delta t\right), i=1,\ldots,n,\ k=1,\ldots,m\right\}. \qquad (1.3)$$

Отбор информативных признаков: периодограмма (1.4), логарифм периодограммы (1.5).

$$X_i\left(q\right)=\frac{1}{m}\left|\sum\nolimits_{k=1}^{m} x_i\left(k\right)e^{-j2\pi qk\frac{1}{m}}\right|^2 \qquad (1.4)$$

$$\widetilde{X}_i\left(q\right)=20\log_{10} X_i\left(q\right) \qquad (1.5)$$

Сокращение размерности полученного вектора до первых значимых *d* компонентов логарифма периодограммы. В результате получаем набор входных данных:

$$y_i=\left[\widetilde{X}_i\left(1\right),\widetilde{X}_i\left(2\right),\ldots,\widetilde{X}_i\left(d\right)\right] \qquad (1.6)$$

Выбор метрики (1.7) на основе признаков, отобранных на шаге 1. Эффективность использования в кластерных вычислениях логарифмической метрики, основанной на периодограмме, обоснована в работе [21].

$$d_{LP}\left(x_i,x_j\right)=\sqrt{\sum\nolimits_{q=1}^{\left\lfloor\frac{m}{2}\right\rfloor}\left[\tilde{X}_i\left(q\right)-\tilde{X}_j\left(q\right)\right]^2} \qquad (1.7)$$

1.2.2. Описание процедур кластеризации

Существует достаточно большое количество процедур, предназначенных для обнаружения аномалий [22]. Метод опорных векторов [23] и метод изолирующего леса [24], которые уже стали своего рода универсальными и наиболее распространенными для решения различных задач обнаружения аномалий. Тем не менее, методы, учитывающие специфику конкретной задачи, обычно показывают лучшие результаты. Для рассматриваемой нами задачи диагностики оборудования целесообразно использовать методы кластеризации,

т. к. можно предположить, что данные, характерные для стабильной работы, образуют четкие кластеры. В этом случае кластеризация позволит выделить нетипичные объекты, которые не подходят ни к одному из кластеров.

Для проведения сравнительного анализа выбраны следующие методы: метод опорных векторов [23, 25] и метод К-средних [26, 27], являющийся наиболее широко распространенным методом кластеризации.

1.2.2.1. Метод опорных векторов

Метод опорных векторов — алгоритм обучения с учителем для решения задач классификации и регрессии, относится к классу линейных классификаторов. Метод опорных векторов, рассматривающийся здесь, является обобщением линейного классификатора на случай построения нелинейных разделяющих плоскостей и основан на прямой задаче оптимизации:

$$\begin{cases} \frac{1}{2}w^T w + C\sum_{i=1}^{2N}\xi_i \to \min\limits_{w,b,\xi} \\ y_i\left(w^T\varphi\left(x_i\right)+b\right) \ge 1-\xi_i \quad ,i=1,\ldots,2N, \\ \xi_i \ge 0 \end{cases} \quad (1.8)$$

где нелинейное преобразование $\varphi\left(x_i\right)$ отображает x_i в пространство информативных признаков с большей размерностью, а C — параметр регуляризации.

Вектор w — вектор, нормальный к оптимальной разделяющей гиперплоскости в преобразованном пространстве, ξ_i — переменные, измеряющие величину ошибки на объектах x_i. Классифицирующая функция:

$$sgn\left(w^T\varphi\left(x\right)+b\right). \quad (1.9)$$

Решение задачи с ограничениями (1.8) получено с помощью метода множителей Лагранжа, в частности, решение двойственной задачи в форме:

$$\begin{cases} \left(\frac{1}{2}\alpha^T Q\alpha\right) - \sum_{i=1}^{2N}\alpha_i \to \min\limits_{\alpha} \\ \sum_{i=1}^{2N} y_i\alpha_i = 0, \qquad i=1,\ldots,2N, \\ 0 \le \alpha_i \le C \end{cases} \quad (1.10)$$

где $\alpha \in R^{2N}$ — вектор множителей Лагранжа, Q — положительно полуопределенная матрица размерности $2N\times 2N$: $Q_{i.j} = y_i y_j k\left(x_i x_j\right)$, k — функция ядра, представляющая скалярное произведение в преобразованном пространстве высокой размерности.

Преимущество использования двойственной формы заключается в том, что вместо вычисления преобразования $\varphi\left(x_i\right)$ требуется только вычисление ядра, и все члены, содержащие множители ξ_i, обнуляются, остается лишь константа C как дополнительное ограничение на множители Лагранжа.

После того, как задача (1.10) решена, только несколько переменных $\alpha_i \ne 0$ и соответствующие им x_i будут являться опорными векторами. С использованием прямых-двойственных отношений вектор w, нормальный к оптимальной разделяющей гиперплоскости и смещение b вычисляются как:

$$w = \sum_{i=1}^{2N} y_i\alpha_i\varphi\left(x_i\right) \quad (1.11)$$

$$b = \frac{1}{N_{SV}}\sum_{i\in SV} w^T\varphi\left(x_i\right) - y_i, \quad (1.12)$$

где $SV \equiv \{i : \alpha_i > 0\}$ — набор индексов опорных векторов, N_{SV} — количество опорных векторов.

Скалярное произведение $w^T\varphi(x_i)$ необходимо для вычисления смещения, и теперь классифицирующая функция (1.9) вычисляется с использованием ядра:

$$\begin{aligned} w^T\varphi\left(x_i\right) &= \sum_{i=1}^{2N} y_i\alpha_i\varphi\left(x_i\right)^T, \\ \varphi\left(x_j\right) &= \sum_{i=1}^{2N} y_i\alpha_i k\left(x_i, x_j\right) \end{aligned} \quad (1.13)$$

Таким образом, непосредственное вычисление функции $\varphi\left(x_i\right)$ не требуется, и функция классификации (1.9) приобретает вид:

$$sgn\left(\sum_{i=1}^{2N} y_i\alpha_i k\left(x_i, x_j\right) + b\right) \quad (1.14)$$

Решение задачи (1.10) осуществляется средствами последовательной минимальной оптимизации. Представленный метод опорных векторов требует выбора функции ядра k и параметра C.

1.2.2.2. Метод кластеризации К-средних

Набор данных y_i должен быть разбит на c кластеров $\{S_1, S_2, \ldots, S_c\}$ таким образом, что-

бы минимизировать сумму квадратов для каждого кластера:

$$\sum_{i=1}^{c}\sum_{y_j\in S_i}\left\|y_j-y_i\right\|^2. \qquad (1.15)$$

Метод K-средних определяет эвристическую стратегию, которая использует итеративный метод группировки, чтобы найти значение (1.15). После начального распределения объектов по кластерам, алгоритм итеративно повторяет два шага:

1. E-шаг. Присваивание каждого вектора-точки y_i кластеру S_i с ближайшим вектором средних:

$$S_i=\left\{\begin{matrix} y_j:\left\|y_j-v_i\right\|^2\le\left\|y_j-v_k\right\|^2, \\ k=1,\ldots,c \end{matrix}\right\}. \qquad (1.16)$$

2. M-шаг. Вычисление новых векторов центровкой точек в каждом кластере:

$$v_i=\frac{1}{\left|S_i\right|}\sum_{y_j\in S_i}y_j. \qquad (1.17)$$

Алгоритм достигает локального минимума функции (1.15), когда прекращается присваивание. Для первоначального распределения центры кластеров часто выбираются случайно. Следует отметить, что не существует гарантий, что алгоритм достигнет глобального минимума в результате работы, и результат может зависеть от начального распределения. Быстрое выполнение алгоритма позволяет осуществить несколько вычислений для различных начальных условий и выбрать лучший результат.

Для выбора числа кластеров с используется мера внутренней валидности, определенная индексом Данна [28]. Для набора кластеров индекс Данна вычисляется как процент между минимальным расстоянием внутри кластеров и максимальным размером кластеров. Для вычисления расстояния внутри кластера и размера кластера существует множество методов, например:

$$\Delta_i=\frac{1}{\left|S_i\right|}\sum_{y_j\in S_i}\left\|y_j-v_i\right\| \qquad (1.18)$$

$$\delta\left(S_i,S_j\right)=\left\|v_i-v_j\right\| \qquad (1.19)$$

Индекс Данна вычисляется как:

$$DI_c=\frac{\min\limits_{k\ne j}\delta\left(S_k,S_j\right)}{\max\limits_{i=1\ldots c}\Delta_i} \qquad (1.20)$$

Оптимальное количество кластеров будет соответствовать большему значению индекса Данна.

2. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ И ИХ ОБСУЖДЕНИЕ

2.1. Выбор критериев для сравнения

Для проведения анализа работы процедур доступно шесть наборов тестовых данных. Для оценки точности диагностики метод опорных векторов применялся к пяти наборам тестовых данных, набор данных № 6 использовался в качестве обучающего. В остальных случаях каждая процедура применялась к шести наборам тестовых данных. Для обучения использовались сбалансированные классы, что позволяет в качестве функционала качества классификации использовать долю правильных ответов алгоритма, вычисляемую по формуле:

$$A=\frac{P}{N}, \qquad (1.21)$$

где N — общее количество объектов, P — количество корректно идентифицированных объектов.

Для оценки точности классификации методом опорных векторов используется метод скользящего контроля по K блокам (1.22). Для проведения эксперимента возьмем $K=5$.

$$CV_K=\frac{1}{K}\sum_{i=1}^{K}Q(\mu\left(\frac{T^l}{F_i}\right),F_i) \qquad (1.22)$$

$$T^l=\bigcup_{i=1}^{K}F_i, \qquad (1.23)$$

где T — обучающая выборка, Q — функционал качества, в данном случае вычисляющийся по формуле (1.21), K — количество блоков.

Для кластеризации методом K-средних вычисление функционала качества (1.21) возможно с использованием контрольных данных значений освещенности.

2.3. Организация вычислительного эксперимента

Для проведения вычислительных экспериментов использовались данные силы тока

и напряжения, полученные в результате реализации симулятора системы электрификации поезда, описанного в работе [29]. Симулятор моделирует работу системы электрификации переменного тока промышленной частоты 50 Гц 2×25 кВ, учитывает факторы нагрузки и возможных отказов системы электрификации и предназначен для изучения тяговой сети электроподвижных составов (ЭПС). Запись показаний осуществляется с частотой 20 кГц.

Для проведения вычислительных экспериментов выполнена программная реализация рассматриваемых методов кластеризации на языке Java.

2.3. Настройка и результаты работы процедур

Для получения оптимальных параметров обработки входных данных d и m осуществляется выполнение процедуры для набора возможных значений параметров и выбираются значения, для которых алгоритмы показали наилучшие результаты.

Параметры обработки данных и алгоритма метода опорных векторов (параметры C и γ выбираются методом поиска по сетке с экспоненциальным ростом частоты):

1. В качестве ядра используется гауссова радиальная базисная функция:

$$k_{RBF} = \left(x_i, x_j\right) = e^{-\gamma\left\|x_i - x_j\right\|^2} \qquad (1.24)$$

2. Временное окно $m = 150$.

3. Количество значимых компонент периодограммы $d = 12$.

4. Параметр $C = 4871$.

5. Параметр $\gamma = 8.6317\times10^{-5}$

Результаты работы метода опорных векторов для задачи обнаружения аномалий приведены в табл. 1.

При использовании метода K-средних были выбраны следующие параметры обработки данных: количество кластеров $c=4$; временное окно $m = 160$; количество значимых компонент периодограммы $d = 10$. Результаты работы алгоритма кластеризации методом K-средних приведены в табл. 2.

Таблица. 1. Результаты работы метода опорных векторов
[Table 1. Results of Support Vector Machine execution]

№ набора данных	Максимальная оценка скользящего контроля, %	Доля правильных ответов, % алгоритма, обученного на наборе данных № 6
1	91.0682	74.7682
2	92.1136	78.2414
3	92.2045	83.4477
4	94.1932	76.1932
5	89.7273	84.5523
6	90.2160	–

Результаты сравнения работы алгоритмов для выявления электрических дуг длительностью более 5 мс приведены на рис. 2.

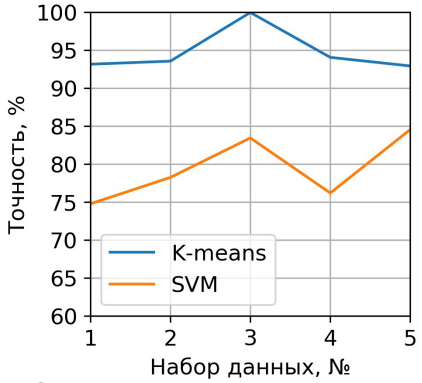

Рис. 2. Сравнение точности алгоритмов
[Fig. 2. Accuracy comparison of algorithms]

Были подобраны оптимальные параметры обработки входных значений и параметры алгоритма для получения наиболее точной оценки аномальности значений для каждой процедуры. Для процедуры кластеризации методом опорных векторов средняя точность диагностики при оптимальных параметрах равна 85.0523 %, а для процедуры кластеризации методом K-средних — 94.7435 %, что характеризует достаточно точное обнаружение аномальных значений и показывает превосходство метода K-средних в контексте решения поставленной задачи. В дальнейшем плани-

Таблица 2. Результаты работы алгоритма кластеризации K-средних
[Table 2. Results of K-means clustering algorithm execution]

Опыт	Кластер	Мощность кластера	Среднее сигнала фотосенсора (освещенность, Вт/см^2)	Среднеквадратичное отклонение сигнала фотосенсора (освещенность, Вт/см^2)
1	1	9042	0.0347	0.0281
	2	12,624	0.0197	0.0184
	3	4935	0.0161	0.0008
	4	1830	0.0167	0.0007
2	1	8679	0.0490	0.0345
	2	12,848	0.0251	0.0229
	3	4106	0.0215	0.0015
	4	1303	0.0214	0.0007
3	1	8672	0.0079	0.0063
	2	12,550	0.0050	0.0019
	3	6761	0.0047	0.0013
	4	1176	0.0047	0.0008
4	1	8573	0.0522	0.0444
	2	14,614	0.0385	0.0150
	3	10,594	0.0350	0.0051
	4	599	0.0376	0.0050
5	1	9295	0.0281	0.0211
	2	13,392	0.0201	0.0089
	3	4341	0.0162	0.0037
	4	828	0.0076	0.0009
6	1	8662	0.0543	0.0507
	2	13,313	0.0315	0.0211
	3	4929	0.0292	0.0015
	4	617	0.0294	0.0011

руется рассмотреть возможность применения алгоритма кластеризации методом K-средних в ансамбле с методами из групп машинного обучения с целью увеличения точности диагностики тяговой сети ЭПС переменного тока.

ЗАКЛЮЧЕНИЕ

Для проактивной стратегии обслуживания были рассмотрены основные компоненты, а также преимущества и недостатки внедрения на примере железнодорожной отрасли. Стоит отметить, что с современным развитием высоких технологий проактивная стратегия обслуживания становится стандартом в ключевых индустриальных областях, и разработка систем, реализующих данную стратегию, является актуальным направлением исследований. Задача реализации системы проактивной стратегии обслуживания ставит множество других задач, таких, как: задачи прогнозирования и обнаружения аномалий в работе оборудования, задачи разработки интеллектуальных систем и систем принятия решений, задачи сбора и обработки информации, например, с использованием сенсорных сетей.

В данной статье был проведен сравнительный анализ процедур кластеризации решения задачи обнаружения аномалий в работе системы электрификации ЭПС, решение которой является основной целью разработки системы диагностики контактной сети электрических железных дорог.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. Predictive Maintenance 4.0. – Pricewaterhouse Coopers B.V. and Mainnovation, 2017.

2. Railway Technical Strategy Europe 2019. – Paris : UIC Communications Department ETF, 2019. – 24 p.

3. *Безуглов, А.* Ключевые показатели эффективности при проведении технического обслуживания и ремонта оборудования / А. Безуглов, О. Кислицына // Вопросы инновационной экономики. – 2019. – Т. 9, № 4. – С. 1501–1514. DOI: 10.18334/vinec.9.4.41208.

4. *Fraga-Lamas, P.* Towards the Internet of Smart Trains: A Review on Industrial IoT Connected Railway / P. Fraga-Lamas, T. Fernández-Caramés, L. Castedo // Sensors. – Bazel, 2017. – V. 17, A. 1457. DOI: 10.3390/s17061457

5. *Brahimi, M.* Development of a prognostics and health management system for the railway infrastructure — Review and methodology / M. Brahimi, K. Medjaher, M. Leouatni, N. Zerhouni // 2016 Prognostics and System Health Management Conference. – Chengdu, 2016. – P. 1–8. DOI: 10.1109/PHM.2016.7819783

6. *Kreyszig, E.* Advanced Engineering Mathematics / E. Kreyszig. – 10th edition. – USA : John Wiley & Sons Inc, 2018. – 1280 p. DOI: 10.1002/bimj.19650070232

7. *Sheriff, M. Z.* Improved Fault Detection and Process Safety Using Multiscale Shewhart Charts / M. Z. Sheriff // Journal of Chemical Engineering & Process Technology. – 2017. – V. 2, No. 18. – P. 1–16. DOI: 10.4172/2157-7048.1000328.

8. *Johnson, M. J.* Bayesian Nonparametric Hidden Semi-Markov Models / M. J. Johnson, A. S. Willsky // Journal of Machine Learning Research. – 2013. – V. 1, No. 14. – P. 673–701.

9. *Aggarwal, C. C.* Outlier Analysis / C. C. Aggarwal. – 2d edition. – New York: Springer International Publishing, 2013. – 466 p. DOI: 10.1007/978-1-4614-6396-2

10. *Schölkopf, B.* Estimating the Support of a High-dimensional Distribution / B. Schölkopf [et al] // Neural Computation. – 2001. – V. 7, No. 13. – P. 1443–1471. DOI: 10.1162/089976601750264965

11. *Rousseeuw, P. J.* A Fast Algorithm for the Minimum Covariance Determinant Estimator / P. J. Rousseeuw, K. Van Driessen // Technometrics. – 1999. – V. 3, No 41. – P. 212–223. DOI: 10.1080/00401706.1999.10485670

12. *Головина, А. М.* Выявление аномалий в работе механизмов методами машинного обучения / А. М. Головина, А. Г. Дьяконов // Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных», Москва, Россия, 10–13 октября 2017 г. – Москва : МГУ, 2017. – С. 389–396.

13. *Aggarwal, C. C.* Theoretical foundations and algorithms for outlier ensembles / C. C. Aggarwal, S. Sathe // ACM SIGKDD Explorations Newsletter. – 2015. – V. 1, No. 17. – P. 24–47. DOI: 10.1145/2830544.2830549

14. *Aydin, I.* Fuzzy integral-based multisensor fusion for arc detection in the pantograph-catenary system / I. Aydin, S. Celebi, S. Barmada, M. Tucci // Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit. – 2016. – V. 232. – P. 159–170. DOI: 10.1177/0954409716662090

15. *Li, K.* Arc fault detection based on cluster analysis and electromagnetic radiation / K. Li, Z. Chen, Y. Z. Zhang, Y. Wang, et al. // Electric Machines and Control. – 2018. – V. 22. – P. 94–101. DOI: 10.15938/j.emc.2018.05.012

16. *Aydin, I.* Particle Swarm Based Arc Detection on Time Series in Pantograph-Catenary System / I. Aydin, O. Yaman, M. Karakose, S. Celebi // NISTA 2014 - IEEE International Symposium on Innovations in Intelligent Systems

and Applications, Proceedings. – 2014. – P. 344–349. DOI: 10.1109/INISTA.2014.6873642

17. *Huang, S.* Cluster Analysis Based Arc Detection in Pantograph-Catenary System / S. Huang, L. Yu, F. Zhang, W. Zhu, Q. Guo // Journal of Advanced Transportation. – 2018. – V. 5. – P. 1-12. DOI: 10.1155/2018/1329265

18. *Назаров Н. С.* Особенности спектрального анализа тяговых токов электроподвижного состава железных дорог / Н. С. Назаров, О. Н. Назаров // Современные проблемы совершенствования работы железнодорожного транспорта: межвуз. сб. тр. – Москва : РОАТ, 2013. – С. 88–100.

19. *Бровкин, В. Е.* Анализ видов искажения напряжения на контактной сети переменного тока 25 кВ / В. Е. Бровкин // Студенческий : электрон. научн. журн. – 2019. – V. 50, No. 6. – Режим доступа: https://sibac.info/journal/student/50/132759 (дата обращения: 29.04.2020).

20. *Madisetti, V.* Handbook of Digital Signal Processing / V. Madisetti, D. Williams. — Boka Raton : CRC Press, 1999. – 1760 p. DOI: 10.1080/00401706.1994.10485863

21. *Caiado, J.* A periodogram-based metric for time series classification / J. Caiado, N. Crato, D. Peña // Computational Statistics & Data Analysis. – 2006. – V. 50. – P. 2668–2684. DOI: 10.1016/j.csda.2005.04.012

22. *Chandola, V.* Anomaly Detection: A Survey / V. Chandola, A. Banerjee, V. Kumar // ACM Computing Surveys. – 2009. – V. 41, No. 3. – A. 15, 58 p. DOI: 0.1145/1541880.1541882

23. *Burges, C. J. C.* A tutorial on support vector machines for pattern recognition / C. J. C. Burges // Data Mining and Knowledge Discovery. – 1998. – V. 2, No. 2. – P. 121167.

24. *Liu, F. T.* Isolation Forest / F. T. Liu, T. K. M. Tony, Z. H. Zhou // Proc. of the 2008 Eighth IEEE Int. Conf. on Data Mining. – 2008. – P. 413–422. DOI: 10.1109/ICDM.2008.17

25. *Dr. Andrew, Ng.* Support Vector Machines / Dr. Ng. Andrew. – Stanford : CS229, Machine Learning, Lecture Notes. – 2012. DOI: 10.1007/978-981-15-2770-8_8

26. *Воронцов, К. В.* Метрические алгоритмы классификации и математические методы обучения по прецедентам [Электронный ресурс] // К. В. Воронцов. – Режим доступа: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML1.pdf.

27. *Jain A.* Data Clustering: A Review / A. Jain, M. Murty, P. Flynn // ACM Computing Surveys. – 1999. – V. 31, No. 3. – P. 264–323.

28. *Bezdek, J. C.* Cluster Validation with generalized Dunn's indices / J. C. Bezdek, N. R. Pal // Proc. 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems. – 1995. – P. 190–193. DOI: 10.1109/ANNES.1995.499469

29. *Shenoy, U. J.* MATLAB/PSB based modeling and simulation of 25 kV AC railway traction system — a particular reference to loading and fault conditions / U. J. Shenoy, K. G. Sheshadri, K. Parthasarathy, H. P. Khincha, // TENCON 2004. 2004 IEEE Region 10 Conference. – 2004. – V. 3. – P. 508–511. DOI: 10.1109/TENCON.2004.1414819

Моисеева Татьяна Александровна – магистр по направлению «Математическое обеспечение и администрирование информационных систем».
E-mail: lina.inverse1995@mail.ru
ORCID iD: https://orcid.org/0000-0002-8127-7268

Леденева Татьяна Михайловна – д-р техн. наук, проф., зав. кафедрой вычислительной математики и прикладных информационных технологий Воронежского государственного университета. E-mail: ledeneva-tm@yandex.ru
ORCID iD: https://orcid.org/0000-0002-3944-2266

DOI: https://doi.org/10.17308/sait.2020.2/2921 ISSN 1995-5499
Received 03.05.2020
Accepted 15.06.2020

A COMPARATIVE ANALYSIS OF CLUSTERING TECHNIQUES FOR THE DETECTION OF ANOMALIES IN THE PARAMETERS CHARACTERISING THE FUNCTIONING OF AN ELECTRICAL RAILWAY SYSTEM

© 2020 T. A. Moiseeva✉, T. M. Ledeneva

Voronezh State University
1, Universitetskaya square, 394018 Voronezh, Russian Federation

Annotation. This article considers the problem of anomaly detection and the solutions to this problem by means of clustering techniques. The study was performed on railway equipment. The article describes several maintenance strategies with the predictive maintenance strategy considered to be the most promising. The main components and aims of a predictive maintenance system are discussed with regard to the railway area. The article considers in detail the implementation of the diagnostics module and formulates the anomaly detection problem based on the sample measurements of the electrical parameters of the functioning of a wired railway system. Time-series values of electrical signals are considered to be the input data for the anomaly detection problem. To preprocess the data spectral analysis techniques were used: the estimation of the power spectrum density together with a periodogram-based metric. The data was processed using the sliding window approach. The article presents the results of the comparison of support vector machines and K-means clustering when applied to the test data and evaluates the ratio of correct answers. The optimal parameters were determined.
Keywords: anomaly detection, time series processing, clustering, support vector machine, predictive maintenance.

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. Predictive Maintenance 4.0. Pricewaterhouse Coopers B. V. and Mainnovation, 2017.

2. Railway Technical Strategy Europe 2019. Paris, UIC Communications Department ETF, 2019. 24 p.

3. *Bezuglov A. E., Kislitsyna O. A.* Key performance indicators for equipment maintenance and repair. Russian Journal of Innovation Economics. 2018. V. 20, No. 3. P. 82–89. DOI: 10.18334/vinec.9.4.41208.

4. *Fraga-Lamas P., Fernández-Caramés T., Castedo L.* Towards the Internet of Smart Trains: A Review on Industrial IoT Connected Railway. Sensors. Bazel, 2017. V. 17, A. 1457. DOI: 10.3390/s17061457

5. *Brahimi M., Medjaher K., Leouatni M., Zerhouni N.* Development of a prognostics and health management system for the railway infrastructure — Review and methodology. Prognostics and System Health Management Conference (PHM-Chengdu). Chengdu, 2016. P. 1–8. DOI: 10.1109/PHM.2016.7819783

6. *Kreyszig E.* Advanced Engineering Mathematics. 10th edition. John Wiley & Sons Inc, 2018. 1280 p. DOI: 10.1002/bimj.19650070232

7. *Sheriff M. Z.* Improved Fault Detection and Process Safety Using Multiscale Shewhart Charts. Journal of Chemical Engineering & Process Technology. 2017. V. 2, No. 18. P. 1–16. DOI: 10.4172/2157-7048.1000328

✉ Moiseeva Tatiana A.
e-mail: lina.inverse1995@mail.ru

8. *Johnson M. J, Willsky A. S.* Bayesian Nonparametric Hidden Semi-Markov Models. Journal of Machine Learning Research. 2013. V. 1, No 14. P. 673–701. Available at: http://www.jmlr.org/papers/volume14/johnson13a/johnson13a.pdf

9. *Aggarwal C. C.* Outlier Analysis. 2d edition. New York, Springer International Publishing, 2013. 466 p. DOI: 10.1007/978-1-4614-6396-2

10. *Schölkopf B. [et al]* Estimating the Support of a High-dimensional Distribution. Neural Computation. 2001. V. 7, No 13. P. 1443–1471. DOI: 10.1162/089976601750264965

11. *Rousseeuw P. J., Van Driessen K.* A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics. 1999. V. 3, No. 41. P. 212–223. DOI: 10.1080/00401706.1999.10485670

12. *Golovina A. M., D'yakonov A. G.* Anomaly Detection in Mechanisms Using Machine Learning. Proceedings of the 19th International Conference on Data Analytics and Management in Data Intensive Domains. Moscow State University, Moscow, Russia, October 10–13, 2017. P. 389–396. Available at: http://ceur-ws.org/Vol-2022/paper59.pdf

13. *Aggarwal C. C., Sathe S.* Theoretical foundations and algorithms for outlier ensembles. ACM SIGKDD Explorations Newsletter. 2015. V. 1, No 17. P. 24–47. DOI: 10.1145/2830544.2830549

14. *Aydin I, Celebi S., Barmada S., Tucci M.* Fuzzy integral-based multi-sensor fusion for arc detection in the pantograph-catenary system. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit. 2016. V. 232. P. 159–170. DOI: 10.1177/0954409716662090

15. *Li K., Chen Z., Zhang Y. Z., Wang Y. [et al]* Arc fault detection based on cluster analysis and electromagnetic radiation. Electric Machines and Control. 2018. V. 22. P. 94–101. DOI: 10.15938/j.emc.2018.05.012

16. *Aydin I., Yaman O., Karakose M., Celebi S.* Particle Swarm Based Arc Detection on Time Series in Pantograph-Catenary System. NISTA 2014 – IEEE International Symposium on Innovations in Intelligent Systems and Applications, Proceedings. 2014. P. 344–349. DOI: 10.1109/INISTA.2014.6873642

17. *Huang S., Yu L., Zhang F., Zhu W., Guo Q.* Cluster Analysis Based Arc Detection in Pantograph-Catenary System. Journal of Advanced Transportation. 2018. V. 5. P. 1–12. DOI: 10.1155/2018/1329265

18. *Nazarov N. S., Nazarov O. N.* Particular digital methods of spectrum analysis of railway rolling stock traction current. Actual issues of railway transport work improvements: interuniversity collection of academic works. Moscow, RUT (MIIT), 2013. P. 88-100 available at

19. *Brovkin V. E.* Analysis of types of voltage signal corruptions in catenary system AC 25 kV. Student's : online science journal. 2019. V. 50, No. 6. Available at: https://sibac.info/journal/student/50/132759 (accessed: 29.04.2020).

20. *Madisetti V., Williams D.* Handbook of Digital Signal Processing. Boka Raton, CRC Press, 1999. 1760 p. DOI: 10.1080/00401706.1994.10485863

21. *Caiado J., Crato N., Peña D.* A periodogram-based metric for time series classification. Computational Statistics & Data Analysis. 2006. V. 50. P. 2668–2684. DOI: 10.1016/j.csda.2005.04.012

22. *Chandola V, Banerjee A., Kumar V.* Anomaly Detection: A Survey. ACM Computing Surveys. 2009. V. 41, No. 3, A. 15. P 8. DOI: 0.1145/1541880.1541882

23. *Burges C. J. C.* A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 1998. V. 2, No. 2. P. 121–167. DOI: https://doi.org/10.1023/A:1009715923555

24. *Liu F. T., Tony T. K. M., Zhou Z. H.* Isolation Forest. Proc. of the 2008 Eighth IEEE Int. Conf. on Data Mining. 2008. P. 413-422. DOI: 10.1109/ICDM.2008.17

25. *Dr. Andrew Ng.* Support Vector Machines. Stanford, CS229, Machine Learning, Lecture Notes. 2012. DOI: 10.1007/978-981-15-2770-8_8.

26. *Vorontsov K. V.* Metrical classification algorithms and mathematical methods of training on precedents. Lecture Notes. Available at: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML1.pdf (accessed: 29.04.2020)

27. *Jain A., Murty M., Flynn P.* Data Clustering: A Review. ACM Computing Surveys. 1999. V. 31, No. 3. P. 264–323. Available at: https://dl.acm.org/doi/10.1145/331499.331504

28. *Bezdek J. C, Pal N. R.* Cluster Validation with generalized Dunn's indices. Proc. 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems. 1995. P. 190–193. DOI: 10.1109/ANNES.1995.499469

29. *Shenoy U. J., Sheshadri K. G., Parthasarathy K., Khincha H. P.* MATLAB/PSB based modeling and simulation of 25 kV AC railway traction system - a particular reference to loading and fault conditions. TENCON 2004. 2004 IEEE Region 10 Conference. 2004. V. 3. P. 508–511. DOI: 10.1109/TENCON.2004.1414819

Moiseeva Tatiana A. – master's degree student of the programme "Mathematical Support and Administration of Information Systems".
E-mail: lina.inverse1995@mail.ru
ORCID iD: https://orcid.org/0000-0002-8127-7268

Ledeneva Tatyana M. – DSc in Technical Sciences, Professor, Head of the Department of Computational Mathematics and Applied Information Technologies, Voronezh State University.
E-mail: ledeneva-tm@yandex.ru
ORCID iD: https://orcid.org/0000-0002-3944-2266