Анализ UMAP – метода снижения размерности исходных данных в машинном обучении для прогнозирования отказов в локомотивном комплексе

Analysis of UMAP, the method for reducing the dimensionality of initial data in machine learning for the purpose of failure prediction in a motive power service

Проневич О.Б.[1], Клокова А.П.[1]*
Pronevich O.B.[1], Klokova A.P.[1]*
[1]*АО «НИИАС», Москва, Российская Федерация*
[1]*JSC NIIAS, Moscow, Russian Federation*
**i@aklokova.ru*

Проневич О.Б.

Клокова А.П.

Резюме. Цель. *Преобразование признаков является одним из этапов применения методов машинного обучения, оказывающих существенное влияние на качество регрессионных моделей. Целью настоящей статьи является разработка критериев оценки качества снижения размерности данных на этапе преобразования признаков и адаптация метода UMAP к решению задачи прогнозирования количества дней до отказа на локомотивах ОАО «РЖД».* **Методы.** *Методы преобразования данных делятся на две группы: одни пытаются сохранить глобальную структуру данных, вторые – локальные расстояния между точками. В настоящей статье подробно рассмотрен нелинейный метод снижения размерности UMAP, низкоразмерное представление данных в котором основывается на преобразовании графа ближайших соседей с сохранением структуры данных. Изучение структуры многообразия исходных данных осуществляется с применением методов топологического анализа данных и методами построения нечетких симплицированных множеств.* **Результаты.** *Анализ теоретической базы UMAP, впервые проведенный на русском языке, позволил обоснованно выделить три основных параметра метода, варьирование которых оказывает существенное влияние на вид данных, полученных в результате преобразования. В частности – на качество разделения классов на двумерном пространстве. Также были определены характеристики входного набора параметров, влияющих на результаты UMAP. Продемонстрированы результаты практического применения метода UMAP. На промежуточных этапах: перечень ближайших соседей, взвешенный граф ближайших соседей. Основной результат: низкоразмерное представление данных (из исходных 44 измерений) на двумерном пространстве с разделением классов, что подтверждается как расчетами, так и визуально.* **Выводы.** *Определено, что UMAP является эффективным и обоснованным методом снижения размерности, позволяющим за счет варьирования параметров преобразовывать данные таким образом, чтобы повысить качество подаваемых на модели машинного обучения данных по критерию «очевидное разделение классов». Преобразование является промежуточным этапом для подготовки данных к применению регрессионных моделей, и разделение классов выполнено для исключения грубых ошибок регрессий.*

Abstract. Aim. *Feature transformation is one of the stages of machine learning application that has a significant effect on the quality of regression models. The paper aims to develop criteria for evaluating the quality of data dimensionality reduction at the stage of feature transformation and adaptation of the UMAP method to the problem of prediction of the number of days to failure in the locomotives of JSC RZD.* **Methods.** *The data transformation methods are divided into two groups, those that attempt to preserve the global data structure, and those that attempt to preserve the distances between points. The paper examines in detail the UMAP no-linear method of dimensionality reduction, whose low-dimensional data presentation is based on a transformation of a nearest neighbour graph retaining the data structure. The structure of the initial data manifold is examined using topological data analysis and simplified fuzzy set construction methods.* **Results.** *The analysis of UMAP theory conducted in the Russian language for the first time enabled a substantiated identification of the three primary parameters of the method, whose variation significantly affects the type of data obtained as the result of a transformation. In particular, that pertains to the quality of class separation over a two-dimensional space. Additionally, the characteristics of the input set of parameters were identified that affect the UMAP results. Practical results of UMAP application were demonstrated. Intermediate results included a list of nearest neighbours, a weighted graph*

of nearest neighbours. The fundamental result is a low-dimensional data representation (out of 44 initial measurements) over a two-dimensional space with class separation, which is confirmed both by calculations, and visually. **Conclusions.** *It was identified that UMAP is an efficient and substantiated method of dimensionality reduction that allows – through parameter variation – transforming data in such a way as to improve the quality of data submitted to machine learning models by the criterion of "evident class separation". The transformation is an intermediate stage of data preparation for regression model application, and class separation was performed for the purpose of eliminating the probability of gross regression errors.*

Ключевые слова: *машинное обучение, снижение размерности, нелинейные методы, преобразование признаков, регрессия, опасные события.*
Keywords: *machine learning, dimension reduction, non-linear methods, feature transformation, regression, hazardous events.*

Для цитирования: *Проневич О.Б., Клокова А.П. Анализ UMAP – метода снижения размерности исходных данных в машинном обучении для прогнозирования отказов в локомотивном комплексе. // Надежность. 2022. №4. С. 53-62. https://doi.org/10.21683/1729-2646-2022-22-4-53-62*
For citation: *Pronevich O.B., Klokova A.P. Analysis of UMAP, the method for reducing the dimensionality of initial data in machine learning for the purpose of failure prediction in a motive power service. Dependability 2022;4:53-62. https://doi.org/10.21683/1729-2646-2022-22-4-53-62*

Поступила *29.08.2022* / **После доработки** *24.10.2022* / **К печати** *15.12.2022*
Received on: *29.08.2022* / **Revised on:** *24.10.2022* / **For printing:** *15.12.2022.*

Введение

Методы машинного обучения успешно применяются для прогнозирования событий на железнодорожном транспорте. В статьях [1-3] рассмотрено решение задач классификации для прогнозирования появления опасного события на объектах пути и электроснабжения железнодорожного транспорта. В этих работах не приводился обзор результатов применения регрессионных моделей, так как с их помощью не удавалось достичь требуемых показателей точности. При моделировании редких событий, когда опасный отказ появляется только в 2% случаев (менее 100 опасных отказов/год на одной железной дороге) методы классификации показывали себя значительно лучше. Однако, используя только методы классификации событий, мы были лишены возможности исследовать такие важные показатели как: количество дней между отказами, вероятность отказа через время Δt. Знание таких показателей позволило бы решать задачи планирования загрузки предприятий, осуществляющих ремонт объектов железнодорожного транспорта, уменьшать время простоя.

Тяговый подвижной состав представляет собой более технически сложное устройство, чем верхнее строение железнодорожного пути (ВСП) и значительная часть объектов железнодорожного электроснабжения, которые были рассмотрены в работах [2-4]. Для классификации событий, связанных с тяговым подвижным составом, разработан обширный «Классификатор опасных отказов технических средств системы КАСАНТ», включающий в себя 469 причин опасных отказов локомотивов и их элементов. Внимание к классификации отказов (нормирование причин и мест отказов) и автоматизированная система сбора информации об отказах – причины, по которым мы обладаем значительными статистическими данными об опасных отказах на локомотивах. На выборке 2019–2021 гг. доля отказов составляет 30%, доля опасных отказов приборов безопасности и радиосвязи локомотива – 10,3%. Может показаться, что 10,3% не так далеки от 2%, однако важны не только относительные, но и количественные оценки. Эти 10,3% являются 4964 отказами. Такой объем данных позволяет применять целый ряд методов, которые мы не могли использовать в случаях, когда количество отказов в обучающей выборке не превышало нескольких сотен. В частности, мы можем применять регрессионные модели, поскольку можем перейти от дискретных значений «был отказ / не было отказа» к условно-непрерывной величине «количество дней до отказа». В табл. 1 и 2 приведены данные о среднем количестве дней между опасными отказами приборов безопасности и радиосвязи локомотива для 8 серий локомотивов (серия локомотивов и названия дорог закодированы). Значения «365» или «Не было отказов» означают, что отказы на локомотивах проходили не чаще, чем один раз в год. Одна из причин этого – малое количество эксплуатируемых локомотивов конкретной серии на конкретной дороге.

Как видно из табл. 1 и 2, значение количества дней между отказами (для приборов безопасности, далее – $m_{t\_\text{прибор}}$) имеет достаточно сильный разброс: от 2,14 до 154,5 и 365 дней. Интересно, что $m_{t\_\text{прибор}}$ заметно отличается на различных дорогах. Например, для серии 123 $m_{t\_\text{прибор}}$ равно 15,5 и 68,6 для дорог 1 и 17 соответственно. Причин такого разброса несколько: различия в объеме эксплуатации локомотивов, их числе на дорогах, интенсивности движения, условиях эксплуатации. Все эти факторы необходимо учесть для того, чтобы прогнозировать количество дней до отказа на конкретном локомотиве. Современные автоматизированные системы

Табл. 1 – Среднее количество дней между отказами в 2019 году

Серия	Железная дорога 1	Железная дорога 17	Железная дорога 24	Железная дорога 58	Железная дорога 94
123	7,41	15,5	Не было отказов	Не было отказов	Не было отказов
134	365,0	9,52	Не было отказов	Не было отказов	Не было отказов
226	365,0	35,5	53,83	60,0	365,0
234	56,75	Не было отказов	365,0	Не было отказов	45,0
240	365,0	85,0	3,48	2,14	365,0
510	40,16	365,0	42,5	81,0	365,0
530	7,04	365,0	Не было отказов	35,0	Не было отказов
640	52,83	42,85	67,66	Не было отказов	365,0

Табл. 2 – Среднее количество дней между отказами в 2020 году

Серия	Железная дорога 1	Железная дорога 17	Железная дорога 24	Железная дорога 58	Железная дорога 94
123	10,78	68,6	Не было отказов	Не было отказов	Не было отказов
134	365,0	10,34	Не было отказов	Не было отказов	Не было отказов
226	365,0	150,0	365,0	83,0	365,0
234	365,0	Не было отказов	365,0	Не было отказов	39,0
240	365,0	48,5	5,91	1,99	365,0
510	365,0	365,0	154,5	365,0	365,0
530	13,84	365,0	Не было отказов	35,66	Не было отказов
640	42,2	23,57	48,25	Не было отказов	365,0

диагностирования состояния локомотивов обеспечивают нас данными, необходимыми для решения такой сложной задачи, как разработка модели прогнозирования количества дней до отказа (опасного отказа) на основе множества частных характеристик локомотивов.

При построении моделей машинного обучения (ML) для железнодорожного транспорта, исследователи сталкиваются с такими трудностями, как изменение систем учета в автоматизированных системах управления (АСУ), ошибки данных из-за ручного ввода, отсутствие идентификаторов объектов. Для оценки возможности использования автоматизированной системы управления как источника данных для моделей ML, проводятся отдельные исследования [5].

В Дирекции тяги – филиале ОАО «РЖД» более 5 лет разрабатывается и функционирует доверенная среда – один из проектов программы Цифровизации ОАО «РЖД»[1]. В ее основе лежит цифровая платформа для сбора, хранения информации и взаимодействия участников проекта с «озером данных» локомотивного комплекса[1]. Локомотивы в доверенной среде однозначно идентифицируются тремя полями, а значит нет ограничений для использования моделей машинного обучения, описанных в [6].

В табл. 3 приведены сведения о данных, взятых в качестве исходных для разработки моделей прогнозирования количества дней до опасного отказа. На их основе были сгенерированы «синтетические» данные путем обработки категориальных признаков, различных вариаций расчета периодичности событий и оценки интенсивности движения для каждого локомотива.

Табл. 3 – Краткие сведения об исходных данных из доверенной среды локомотивного комплекса

№	Вид информации	Первоначальное количество признаков
1	Отказы	10
2	Замеры колесных пар	18
3	Заводские ремонты локомотивов	46
4	Данные о пробегах, составности и состояниях локомотивов из Автоматизированной системы оперативного управления перевозками	252
5	Данные комиссионных осмотров	12
6	Рекламационная работа	20
7	Данные по замечаниям машиниста из АСУ Замечания машинистов	23
8	Нарушения из АСУ Нарушения безопасности движений	36
9	Статус линейного оборудования	20
10	Отказы линейного оборудования	18
11	Ремонты оборудования	14
12	Результаты расследования нарушений	14
13	Изменения инвентарного парка	79
14	Снятое на ремонт оборудование	12

Более 574 признаков – такой объем данных, с одной стороны, кажется перспективным, т.к. повышает вероятность использования значимых признаков. С другой стороны, для использования в качестве входных параметров модели такого количества признаков требуется

[1] https://gudok.ru/vestnik-cki/?ID=1559003&archive=2021.03.31

задействование больших вычислительных мощностей и большого времени для обучения модели. Последний фактор существенно затрудняет решения задачи подбора параметров модели для достижения требуемых показателей качества. При работе с такими данными необходимо отдельное внимание уделять этапу отбора и преобразования признаков. Основания цель этого этапа – формирование перечня репрезентативных признаков и исключения данных, не имеющих связи с целевой меткой (количество дней до отказа).

Обзор источников

Существует множество этапов преобразования признаков: очистка [7, 8]; генерация новых признаков (как для целей увеличения количества признаков, так и уменьшения) [9]; отбор признаков и снижение размерности [10]. Мы перечислили ключевые этапы, однако для каждой задачи формируется свой перечень применяемых методов.

Отбор признаков, характеризующих состояния локомотивов был осуществлен с применением линейной модели, преобразование осуществлялось методом PowerTransformer[1] (функциональное степенное преобразование, чтобы сделать данные более похожими на гауссовы). Алгоритм отбора и преобразования признаков приведен на рис. 1.

После применения алгоритма исключения отобрано 44 признака, однако даже такое количество может неудовлетворительно сказаться на качестве моделей машинного обучения, и это обостряет вопрос применения дополнительных эффективных методов преобразования признаков. Значительная часть моделей

[1] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PowerTransformer.html

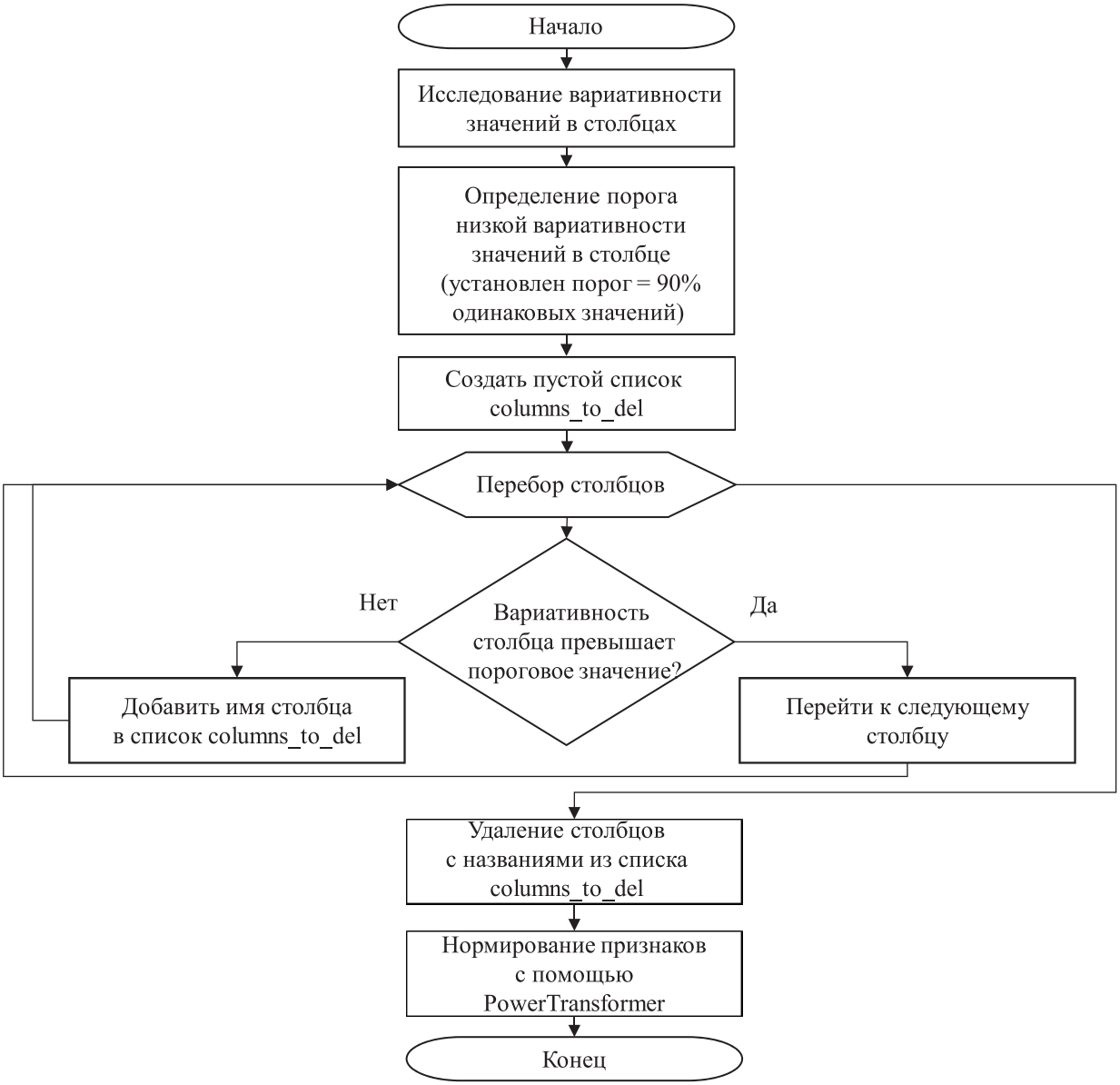

Рис.1. Блок-схема алгоритма исключения неинформативных данных

машинного обучения для железнодорожного транспорта строится для низкоразмерных данных, например, для предиктивного анализа технического состояния систем тепловозов [11], обнаружения неисправностей подшипников качения асинхронного тягового электродвигателя локомотивов [12]. Для прогноза технического состояния газовоздушного тракта тепловозного двигателя – всего 7 признаков [13].

К таким дополнительным методам, способным оказать существенное влияние на качество итоговых моделей, являются методы снижения размерности. Их применение позволяет визуализировать результаты отбора признаков, разработать план по устранению ошибок, допущенных при преобразовании исходных данных.

В работах [14, 15] приведены выводы о применении 6-и методов снижения размерности. Результаты исследования [14] показали преимущество нелинейных методов снижения размерности, в особенности, метод равномерного приближения и проекции (*UMAP* от англ. *uniform approximation and projection*), а также *Isomap* (от англ. *Isometric Mapping)* и многомерное шкалирование (*MDS* от англ. *multidimensional scaling*) для выходных пространств малой размерности (2, 4).

Нелинейные методы являются достаточно сложными и не так часто используются исследователями. Самым популярным, в т.ч. по причине доступности интерпретации результатов, является метод главных компонент *PCA (PCA* от англ. *principal component analysis)* [16]. Таким образом, можно выделить 3 основных метода снижения размерности: *PCA*, метод стохастического вложения соседей с *t*-распределением *(t-SNE* от англ. *t-distributed stochastic neighbor embedding)* и *UMAP*.

Алгоритмы понижения размерности можно разделить на 2 основные группы: они пытаются сохранить либо глобальную структуру данных, либо локальные расстояния между точками. К первым относятся такие алгоритмы как *PCA* и *MDS* , а ко вторым – *t-SNE*, *Isomap*, *LargeVis* (от англ. *large graph visualization*) и другие. *UMAP* относится именно к последним и показывает схожие с *t-SNE* результаты[1]. Ключевые преимущества *UMAP* перед *t-SNE*: скорость вычисления и отсутствие ограничений на размерность исходного пространства признаков, которое необходимо уменьшить.

UMAP является одним из самых молодых методов. Несмотря на то, что он применяется исследователями России с 2018 года[2], в русской литературе мы не встретили подробного разбора математического аппарата данного метода и этапов его применения. Для эффективного использования каких-либо методов (особенно относящихся к классу нелинейных) необходимо понимать их теоретическую основу и алгоритмы работы.

Среди всех вышеописанных методов мы выделяем *UMAP* по следующим причинам:

- успешное применение *UMAP* другими исследователями;
- демонстрируемая авторами метода возможность без учителя снижать размерность таким образом, чтобы области точек, характеризующие объекты различных классов были разнесены в низкоразмерном пространстве [11, 12];
- более высокая скорость вычисления, по сравнению с другими нелинейными методами;
- отсутствие ограничений на размерность исходного пространства признаков, которое необходимо уменьшить.

Ниже приведена краткая история метода, описан математический аппарат, позволяющий эффективно снижать размерность исходных данных. Эффективность снижения размерности исходных данных мы определяем как показатель пересечения классов (аналог ошибки классификации в задачах классификации). В случае задачи прогнозирования количества дней до появления опасного отказа мы можем выделить два класса событий: событие D – наличие опасного отказа в течение 30 дней и событие $\overline{D}$ – отсутствие отказа в течение 30 дней. Для оценки эффективности снижения размерности мы можем использовать те же показатели, что используются для оценки ошибки бинарной классификации. Таким образом, для подбора параметров *UMAP* для поиска оптимального низкоразмерного представления данных мы можем использовать алгоритмы, применяемые при решении задач классификации. Эту задачу невозможно решить без понимания математического аппарата *UMAP*.

Обзор метода *UMAP* с результатами применения для решения задачи снижения размерности при оценке состояния локомотивов

Авторы метода *UMAP* – Лиленд Макиннес, Джон Хили, Джеймс Мелвилл – в 2018 году опубликовали большую статью, включающую в себя описание алгоритмов работы метода, математического аппарата, результатов использования [17]. После апробации на практике в 2020 году появился скорректированный алгоритм [18].

Авторы *UMAP* последовательно решают две крупные задачи: изучение структуры многообразия в многомерном пространстве[2] [18] и поиск низкоразмерного представления данных.

Изучение структуры многообразия осуществляется методами топологического анализа данных и методами построения нечетких симплицированных множеств. В частности, первая задача *UMAP* – построение взвешенного графа k-соседей.

Пусть $X = \{x_1,\ldots,x_N\}$ – входной набор данных с метрикой (или мерой несходства) d: $X \times X \to R \geq 0$. Учитывая входной гиперпараметр k, для каждого x_i мы вычисляем множество $\{x_{i_1},\ldots,x_{i_k}\}$ из k ближайших соседей x_i по

[1] https://habr.com/ru/company/newprolab/blog/350584

[2] https://towardsdatascience.com/umap-dimensionality-reduction-an-incredibly-robust-machine-learning-algorithm-b5acb01de568

метрике d. Это вычисление может быть выполнено с помощью алгоритма поиска любого ближайшего соседа или приближенного алгоритма поиска ближайшего соседа. Авторы *UMAP* рекомендуют использовать алгоритм спуска к ближайшим соседям из [19]. Для каждого x_i определим ρ_i и σ_i [18, с. 14]:

$$\rho_i = \min\left\{d\left(x_i, x_{i_j}\right) | 1 \le j \le k, d\left(x_i, x_{i_j}\right) > 0\right\}, \qquad (1)$$

где ρ_i – расстояние до ближайшего соседа i-го объекта, и установим σ_i равным такому значению, что

$$\sum_{j=1}^{k} \exp\left(\frac{-\max\left(0, d\left(x_i, x_{i_j}\right) - \rho_i\right)}{\sigma_i}\right) = \log_2(k). \qquad (2)$$

Такой подход позволяет соединить x_i по крайней мере с одной другой точкой данных с ребром веса 1, это эквивалентно результирующему нечеткому симплициальному множеству, локально связанному в точке x_i. С практической точки зрения это значительно улучшает представление данных очень высокой размерности, когда другие алгоритмы, такие как *t-SNE*, начинают страдать от проклятия размерности [20]. Выбор σ_i соответствует (сглаженному) нормировочному коэффициенту, определяющему риманову метрику, локальную в точке x_i.

Но, прежде чем говорить о взвешенном графе, остановимся подробнее на определении ближайших соседей. Одной из концепцией *UMAP* является – «сосед моего соседа – мой сосед». При поиске ближайших соседей необходимо определить *n_neighbors* – количество соседей для вычисления графа k-соседей (входной гиперпараметр, задаваемый исследователем, принимает целые значения) и *metric* – метрику для оценки расстояния между наблюдениями. На основании этих параметров строятся графы[1] k-соседей (число соседей равно *n_neighbors*) и осуществляется поиск ближайшего соседа. Выбор метода поиска ближайшего соседа является вопросом отдельного исследования, однако авторы *UMAP* рекомендуют использовать метод, приведенный в работе [19]. Для реализации положений, изложенных в статье Дон Вэйя, Чарикара Мозеса и Кай Ли [19] создана библиотека *PyNNDescent*[4]. В результате на основании исходного массива размера $n \times m$ (n – количество строк, m – количество признаков данных) формируется новый массив размера $n \times n\_neighbors$. Каждая строка содержит в себе индексы *n_neighbors* ближайших соседей, т.е. *PyNNDescent* формирует n наборов данных. При этом одно наблюдение может состоять в разных наборах ближайших соседей. Продемонстрируем на практике работу *PyNNDescent*. Исходный массив данных содержит информацию о состояниях секций локомотивов и имеет размеры: 230475 × 44, где 44 – количество параметров, по которым оценивается локомотив, 230475 – число наблюдений. Пример данных с закодированными названиями столбцов приведен в табл. 4.

[1] https://github.com/lmcinnes/pynndescent

Табл. 4 – Исходный массив данных

Index	X1	X2	X3	X4	…	X42	X43	X44
0	1,0	46,0	0,0	10,0	…	10,0	865,0	3250,0
1	0,0	42,0	0,0	10,0		10,0	861,0	3250,0
2	0,0	10,0	1,0	10,0		10,0	146,0	3080,0
3	0,0	10,0	0,0	10,0		10,0	146,0	3080,0
4	0,0	18,0	0,0	5,0		5,0	316,0	3080,0

Наблюдение с индексом 1 попало в 57 наборов ближайших соседей. Пример соседей наблюдения 1 из набора № 19 приведен в табл. 5.

Табл. 5 – Примеры ближайший соседей в наборе №19

Index	X1	X2	X3	X4	…	X42	X43	X44
Состояния исходного наблюдения								
1	0,0	42,0	0,0	10,0	…	10,0	861,0	3250,0
Соседи								
0	1,0	46,0	0,0	10,0		10,0	865,0	3250,0
17	0,0	42,0	0,0	10,0		10,0	928,0	3080,0

Далее осуществляется построение связующего дерева. Учитывая входной гиперпараметр k, для каждого x_i мы вычисляем множество $\{x_{i_1}, \ldots, x_{i_k}\}$ из k ближайших соседей x_i по метрике d. Метрика d, так же, как и k, задается исследователем. Вопрос выбора оптимальных значений d и k является предметом отдельного исследования и будет рассматриваться в других статьях, продолжающих настоящую работу.

Для этого необходимо определить взвешенный ориентированный граф $G^{-} = (V, E, w)$. Вершины V графа G^{-} образуют множество X. Тогда мы можем сформировать множество направленных ребер $E = \left\{\left(x_i, x_{i_j}\right) | 1 \le j \le k, 1 \le i \le N\right\}$, и определить вес $v_{j|i}$ ребра

$$v_{j|i} = \exp\left(\frac{-\max\left(0, d\left(x_i, x_{i_j}\right) - \rho_i\right)}{\sigma_i}\right), \qquad (3)$$

где $d(x_i, x_{i_j})$ – расстояние между объектами;

x_i – объект наблюдения;

x_{i_j} – j-ый ближайший сосед x_i объекта наблюдения;

ρ_i и σ_i – описаны выше.

При построении взвешенного ориентированного графа, между вершинами могут существовать два ребра с разными весами. Вес ребра интерпретируется как вероятность существования данного ребра, направленного от одного объекта к другому. Исходя из этого, ребра между двумя вершинами объединяются в одно с весом, равным вероятности существования хотя бы одного ребра:

$$w(x_i, x_j) = w(x_i \to x_j) + w(x_j \to x_i) - \\ - w(x_i \to x_j) \cdot w(x_j \to x_i). \qquad (4)$$

Таким образом, алгоритм получает взвешенный неориентированный граф.

На рис. 2 приведен пример такого графа состояний электровоза.

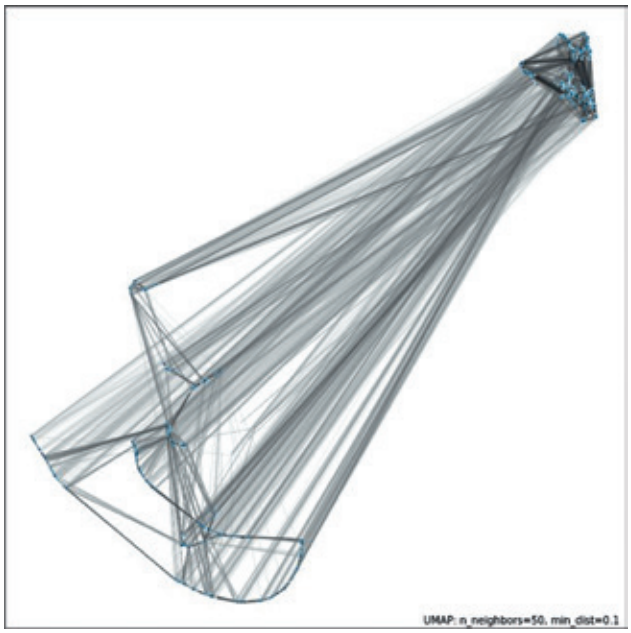

Рис. 2. Взвешенный граф ближайших соседей для электровозов

Шаг 2 – Поиск низкоразмерного представления

Основная задача поиска низкоразмерного представления – вычисление нечеткого топологического представления. В *UMAP* сила притяжения между двумя вершинами i и j с координатами y_i и y_j (точки взвешенного графа H, подробнее ниже) соответственно определяется по формуле:

$$\frac{-2ab\left\|y_i - y_j\right\|_2^{2(b-1)}}{1+\left\|y_i - y_j\right\|_2^2} w((x_i, x_j))(y_i - y_j), \qquad (5)$$

где w_{ij} – низкоразмерное сходство, определяется выражением (6):

$$w_{ij} = \left(1 + a\left\|y_i - y_j\right\|_2^{2b}\right)^{-1}, \qquad (6)$$

где y_i, y_j – координаты вершин i и j; a, b – являются гиперпараметрами, задаваемыми и варьируемыми исследователем. По умолчанию в алгоритме *UMAP* используются следующие начальные значения: $a \approx 1{,}929$ и $b \approx 0{,}7915$. Далее a, b выбираются нелинейным методом наименьших квадратов против кривой Ψ: $R\ d \times R\ d \rightarrow$ [0, 1], где:

$$\Psi\left(y_i, y_j\right) = \begin{cases} 1, & \left|\left\|y_i - y_j\right\|_2 \le min\_dist; \\ exp(-\left\|y_i - y_j\right\|_2 - min\_dist)), \end{cases} \qquad (7)$$

где *min_dist* – минимальное расстояние, на котором точки могут находиться друг от друга в низкоразмерном представлении (значение задается исследователем, по умолчанию 0,1).

Силы отталкивания рассчитываются путем выборки из-за вычислительных ограничений. Таким образом, всякий раз, когда к ребру прикладывается сила притяжения, одна из вершин этого ребра отталкивается выборкой других вершин. Сила отталкивания определяется выражением:

$$\frac{2b}{\left(\epsilon + \left\|y_i - y_j\right\|_2^2\right)\left(1 + a\left\|y_i - y_j\right\|_2^{2b}\right)}(1 - w((x_i, x_j)))(y_i - y_j), \qquad (7)$$

где ϵ – небольшое число для предотвращения деления на ноль (0,001 в текущей реализации).

Силы, описанные выше, получены из градиентов, оптимизирующих перекрестную энтропию по ребрам между взвешенным графом G и эквивалентным взвешенным графом H, построенным из точек $\{y_i\}_i = 1...N$. Цель авторов *UMAP* – расположить точки y_i так, чтобы взвешенный граф, индуцированный этими точками, наиболее близко приближался к графу G, где измеряется разница между взвешенными графами по суммарной кросс-энтропии по всем вероятностям существования ребра. Поскольку взвешенный граф G фиксирует топологию исходных данных, эквивалентный взвешенный граф H, построенный из точек $\{y_i\}_i = 1\ldots N$, соответствует топологии настолько точно, насколько позволяет оптимизация, и, таким образом, обеспечивает хорошее низкоразмерное представление общей топологии данных.

Задача поиска w_{ij} решается путем минимизации функции стоимости C_{UMAP}, которая также является перекрестной энтропией.

$$C_{UMAP} = \sum_{i \ne j} \log\left(\frac{v_{ij}}{w_{ij}}\right) + \left(1 - v_{ij}\right)\log\left(\frac{1 - v_{ij}}{1 - w_{ij}}\right), \qquad (8)$$

где w_{ij} – см. (5); v_{ij} определяется выражением:

$$v_{ij} = \left(v_{j|i} + v_{i|j}\right) - v_{j|i}v_{i|j}, \qquad (9)$$

где $v_{j|i}$ – представляют собой локальное нечеткое симплициальное членство во множестве, основанное на гладких расстояниях ближайших соседей (формула (3)).

Результат представления взвешенного графа ближайших соседей, характеризующего состояния электровозов, на двумерном пространстве приведен на рис. 3. На рис. 3 показаны точки, соответствующие состояниям секций локомотивов: красным цветом – для секций, на которых в течение 30 дней после фиксирования состояния (сбора данных) были отказы, зеленым цветом – для секций, на которых в течение аналогичного периода не было отказов. Числа по осям X и Y не имеют физического смысла, поэтому оси условно обозначены как парам_1 (параметр 1) и парам_2 (параметр 2).

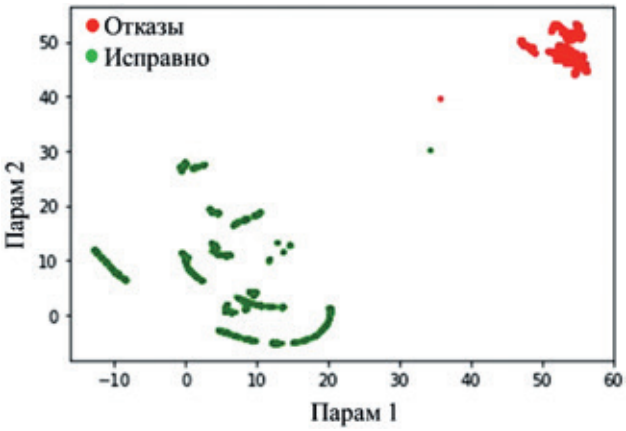

Рис. 3. Низкоразмерное представление данных о состояниях электровозов. Расстояние *jaccard*

Как видно из рис. 3, явно есть наборы точек, однозначно соответствующие состояниям секций локомотивов, предшествующих отказу. Также есть область зеленых – «безопасных» точек. Области зеленых и красных точек имеют пересечения. Всего в выборке 15247 наблюдений. Из них 6158 – «красные» (класс «отказ»), 9089 – «зеленые» (класс «отсутствие отказа»).

На «территории» красных точек оказалось 7 зеленых (меньше 0,1% от общего числа зеленых точек), среди зеленых – 4 красных точки (меньше 0,1%). Таких показателей мы добились путем настройки гиперпараметров *UMAP* и применения методов нормализации исходных данных. Этому исследованию будет посвящена отдельная статья, здесь мы хотим продемонстрировать достигнутые результаты. Анализ математического аппарата позволил нам определить перечень параметров, варьирование которых позволило получить вышеизложенные результаты. В качестве основных параметров выделены:

1) $d(x_i, x_{i_j})$ – метрика оценка расстояния между наблюдениями x_i и x_{i_j};

2) *n_neighbors* – количество соседей;

3) *min_dist* – минимально расстояние, на котором точки могут находиться друг от друга в низкоразмерном представлении;

4) характеристики входного набора параметров:

a. количество признаков;

b. нормирование признаков;

c. баланс классов.

Первые три параметра из списка – важнейшие варьируемые признаки *UMAP*, это указано в документации[1] на метод. Понимание того, каким образом они влияют на итоговый результат, оказывает ключевое влияние на результаты моделирования. В частности, *UMAP* предусматривает возможность использования собственных метрик $d(x_i, x_{i_j})$, не входящих в базовый набор.

По своей сути UMAP относится к классу методов обучения без учителя, однако очевидно, что лучших показателей качества регрессионных моделей мы достигнем, если на вход будем подавать данные, в которых классы максимально разделены. Такой подход позволяет минимизировать число ошибок, связанных с неверной оценкой событий, когда отказа не было.

[1] https://umap.scikit-tda.org/parameters.html

Выводы и перспективы исследования

В рамках настоящей статьи мы рассмотрели математический аппарат метода снижения размерности *UMAP*, основанного на построении графа ближайших соседей и поиске низкоразмерного представления данных – на двумерном пространстве. В ходе преобразования теряется физический смысл осей, однако в новом пространстве сохраняется структура данных (это можно увидеть, соотнеся рис. 2 и 3). В статье впервые на русском языке изложен математический аппарат *UMAP*, определены параметры, варьирование которых позволяет адаптировать метод к решению задачи прогнозирования количеств дней до отказа.

На рис. 3 приведен пример низкоразмерного представления данных о состояниях электровозов, где в качестве метрики расстояния использована метрика *jaccard*, а к исходным данным применялись методы отбора признаков и нормирования. После анализа теоретической базы и первых расчетов определено, что *UMAP* является эффективным методом снижения размерности, т.к. он позволяет «дистанцировать» друг от друга классы событий. Для оценки эффективности снижения размерности исходных данных предложено использовать показатели пересечения классов (аналог ошибки классификации в задачах классификации).

Перспективы исследования для достижения цели повышения качества регрессионных моделей прогнозирования количества дней до отказа:

1) Варьирование параметров *UMAP* для решения задачи разделения локомотивов на классы: «отказ вида *N* / отсутствие отказа вида *N*». Интерпретация результатов и разработка алгоритма подбора параметров для «лучшего» разделения классов;

2) Обучение регрессионных моделей на данных из низкоразмерного представления и разработка алгоритма интерпретации результатов;

3) Обобщение результатов и разработка общего алгоритма построения регрессионных моделей с применением методов разделения классов на этапе отбора и преобразования признаков.

Библиографический список

1. Шубинский И.Б., Проневич О.Б. Методы интеллектуального анализа данных для прогнозирования опасных событий // Железнодорожный транспорт. 2021. № 12. С. 27-31.

2. Проневич О.Б., Зайцев М.В. Интеллектуальные методы повышения точности прогнозирования редких опасных событий на железнодорожном транспорте // Надежность. 2021. Т. 21. № 3. С. 54-65. DOI: 10.21683/1729-2646-2021-21-3-54-65

3. Применение методов машинного обучения для прогнозирования опасных отказов объектов железнодорожного пути / И.Б. Шубинский, А.М. Замышляев, О.Б. Проневич, А.Н. Игнатов, Е.Н. Платонов // Надежность. 2020. Т. 20. № 2. С. 43-53. DOI: 10.21683/1729-2646-2020-20-2-43-53

4. Платонов Е.Н., Просвирин К.В. Прогнозирование дефектов верхнего строения железнодорожного пути методами машинного обучения // Вестник компьютерных и информационных технологий. 2022. Т. 19. № 2. С. 8-18. DOI: 10.14489/vkit.2022.02.pp.008-018

5. Корнеева Е.В. Сидоренко В.Г. Анализ применимости термина Big Data к автоматизированной системе оперативного управления перевозками // Наука и техника транспорта. 2022. № 1. С. 70-76.

6. Устич П.А., Иванов А.А., Мажидов Ф.А. Применение информационных технологий в системе технического обслуживания и ремонта вагонов // Автоматизация. Современные технологии. 2016. № 10. С. 29-38.

7. Калайдин Е.Н., Пиронко М.Д. Особенности сбора и обработки данных для построения моделей машинного обучения // Актуальные проблемы экономической теории и практики: сборник научных трудов / под редакцией В.А. Сидорова. Краснодар, 2020. С. 116-123.

8. Тимченко Е.А. Проблемы предочистки данных // Молодежная наука – развитию агропромышленного комплекса: материалы Всероссийской (национальной) научно-практической конференции студентов, аспирантов и молодых ученых. Курск, 3-4 декабря 2020 г. С. 263-269.

9. Акимов А.А., Валитов Д.Р., Кубряк А.И. Предварительная обработка данных для машинного обучения // Научное обозрение. Технические науки. 2022. № 2. С. 26-31. DOI: 10.17513/srts.1391

10. Анализ существующих методов снижения размерности входных данных / С.Д. Ерохин, Б.Б. Борисенко, И.Д. Мартишин, А.С. Фадеев // Телекоммуникации и транспорт. 2022. Т. 16. № 1. С. 30-37. DOI: 10.36724/2072-8735-2022-16-1-30-37

11. Федотов М.В., Грачев В.В. Предиктивная аналитика технического состояния систем тепловозов с использованием нейросетевых прогнозных моделей // Бюллетень результатов научных исследований. 2021. № 3. С. 102-114. DOI: 10.20295/2223-9987-2021-3-102-114.

12. Хамидов О.Р., Грищенко А.В. Обнаружение неисправностей подшипников качения асинхронного тягового электродвигателя локомотивов на основе современных интеллектуальных методов // Вестник транспорта Поволжья. 2020. № 1 (79). С. 35-41.

13. Диагностирование газовоздушного тракта тепловозного дизеля с использованием интеллектуального классификатора / В.В. Грачев, М.В. Федотов, А.В. Грищенко, Ф.Ю. Базилевский, А.Л. Шарапов // Бюллетень результатов научных исследований. 2022. № 2. С. 124-140. DOI: 10.20295/2223-9987-2022-2-124-140

14. Ефименко Е.Ю., Мясников Е.В. Оценка методов снижения размерности в задаче распознавания личности по походке // Сборник трудов по материалам VII Международной конференции и молодежной школы «Информационные технологии и нанотехнологии» (ИТНТ 2021), Самара, 20-24 сентября 2021. Самарский университет, 2021. Т. 2. С. 159-160.

15. Горбунова А.А. Сравнительный анализ алгоритмов снижения размерности данных для исследования экспрессии генов // 77-я научная конференция студентов и аспирантов Белорусского государственного университета: материалы конференции в 3 ч., Минск, 11–22 мая 2020 года. Минск: Белорусский государственный университет, 2020. С. 161-164.

16. Кулагин М.А. Интеллектуальная система анализа и прогнозирования нарушений при управлении подвижным составом: дис. … канд. техн. наук: 2.9.8. М., 2022. 229 с.

17. Leland McInnes, John Healy, James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Электронный ресурс]. Режим доступа: https://arxiv.org/pdf/1802.03426.pdf, свободный (дата обращения 03.10.2022). DOI: 10.48550/arXiv.1802.03426

18. Leland McInnes, John Healy, James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction // arXiv. — 2020. — 21 September. DOI: https://doi.org/10.48550/arXiv.1802.03426

19. Wei Dong, Charikar Moses, Kai Li. Efficient k-nearest neighbor graph construction for generic similarity measures // Conference: Proceedings of the 20th international conference on World wide web, March 28–April 1, 2011. Hyderabad, India, 2011. Pages 577–586. DOI: 10.1145/1963405.1963487

20. Kai Ming Ting, Takashi Washio, Ye Zhu et al. Breaking the curse of dimensionality with Isolation Kernel [Электронный ресурс]. Режим доступа: https://arxiv.org/pdf/2109.14198.pdf, свободный (дата обращения 03.10.2022). DOI: 10.48550/arXiv.2109.14198

References

1. Shubinsky I.B, Pronevich O.B. [Methods of deep learning for hazard prediction]. *Zheleznorodozhny transport* 2021;12:27-31. (in Russ.)

2. Pronevich O.B., Zaytsev M.V. Intelligent methods for improving the accuracy of prediction of rare hazardous events in railway transportation. *Dependability* 2021;21(3):54-65. DOI: https://doi.org/10.21683/1729-2646-2021-21-3-54-65

3. Shubinsky I.B., Zamyshliaev A.M., Pronevich O.B., Platonov E.N., Ignatov A.N. Application of machine learning methods for predicting hazardous failures of railway track assets. *Dependability* 2020;2:45-53. DOI: https://doi.org/10.21683/1729-2646-2020-20-2-43-53

4. Platonov E.N., Prosvirin K.V. Prediction of track structure defects by machine learning methods. *Herald of computer and information technologies* 2022;19(2):8-18. DOI: 10.14489/vkit.2022.02.pp.008-018 (in Russ.)

5. Korneeva E.V., Sidorenko V.G. Analysis of big data term applicability to automated system of transportation

operational control. *Science and Technology in Transport* 2022;1:70-76. (in Russ.)

6. Ustich P.A., Ivanov A.A., Mazhidov F.A. Application of information technology in the cars technical maintenance system and repair. *Avtomatizatsiya. Sovremennye tekhnologii* 2016;10:29-38. (in Russ.)

7. Kalaydin E.N., Pironko M.D. [Specificity of the collection and processing of data for the purpose of construction of machine learning models]. In: Sidorov V.A., editor. [Topical issues of economic theory and practice. Collected science papers]. Krasnodar; 2020. P. 116-123. (in Russ.)

8. Timchenko E.A. [Matters of preliminary data cleansing]. In: [Young Science for the Development of Agriculture. Proceedings of the All-Russian (National) research and practice conference of undergraduate, postgraduate students and young scientists]; 2020:263-269. (in Russ.)

9. Akimov A.A., Valitov D.R., Kubryak A.I. Data preprocessing for machine learning. *Scientific Review. Technical science* 2022;2: 26-31. DOI: 10.17513/srts.1391 (in Russ.)

10. Erokhin S.D., Borisenko B.B., Martishin I.D., Fadeev A.S. Analysis of existing methods to reduce the dimensionality of input data. *T-Comm* 2022;16(1):30-37. DOI: 10.36724/2072-8735-2022-16-1-30-37 (in Russ.)

11. Fedotov M.V., Grachev V. V. Predictive analytics of the technical condition of diesel locomotive systems using neural network predictive models. *Bulletin of Scientific Research Results* 2021;3:102-114. DOI 10.20295/2223-9987-2021-3-102-114. (in Russ.)

12. Khamidov O.R., Grishchenko A.V. [Detecting faults in rolling bearings of asynchronous traction electric motors of locomotives using modern AI-based methods]. *Vestnik transporta Povolzhya* 2020;1(79):35-41. (in Russ.)

13. Grachev V.V., Fedotov M.V., Grizhshenko A.V., Bazilevskiy F.Yu., Sharapov A.L. Locomotive Diesel Gas-Air Tract Diagnostics with the Use of Intellectual Classifier. *Bulletin of Scientific Research Results* 2022;2:124-140. DOI 10.20295/2223-9987-2022-2-124-140. (in Russ.)

14. Efimenko E.Yu., Miasnikov E.V. [Evaluating the methods of dimensionality reduction as part of identity recognition by the walk]. In: Miasnikov V.V., editor. [Proceedings of the VII International Conference and Youth School]. Samara; 2021. (in Russ.)

15. Gorbunov A.A. [Comparative analysis of the data dimensionality reduction algorithms as part of gene expression research]. In: [Proceedings of the 77-th Science Conference of the Undergraduate and Postgraduate Students of the Belarusian State University in 3 volumes]. Minsk; 2020. P. 161-164. (in Russ.)

16. Kulagin M.A. [An AI-based system for analysing and predicting train control violations: a Candidate of Engineering Thesis]. Moscow; 2022. (in Russ.)

17. McInnes L., Healy J., Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv; 2018. DOI: https://doi.org/10.48550/arXiv.1802.03426

18. McInnes L., Healy J., Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv; 2020. DOI: https://doi.org/10.48550/arXiv.1802.03426

19. Dong W., Moses C., Li K. Efficient k-nearest neighbor graph construction for generic similarity measures. In: Proceedings of the 20th international conference on World wide web; 2011. P. 577-586. DOI: 10.1145/1963405.1963487

20. Ting K.M., Washio T., Zhu Y., Xu Y. Breaking the curse of dimensionality with Isolation Kernel. arXiv; 2021. DOI: https://doi.org/10.48550/arXiv.2109.14198

Сведения об авторах

Проневич Ольга Борисовна – кандидат технических наук, руководитель проектов отдела постановки задач, внедрения и сопровождения системных разработок отделения управления рисками сложных технических систем АО «НИИАС», Москва, Российская Федерация, e-mail: obpronevich@gmail.com

Клокова Анна Павловна – аспирант, Российский университет транспорта РУТ (МИИТ); специалист отделения управления рисками сложных технических систем АО «НИИАС», Москва, Российская Федерация, e-mail: i@aklokova.ru

About the authors

Olga B. Pronevich, Candidate of Engineering, Project Manager, Unit for Problem Definition, Deployment and Support of System-Level Designs, Division for Risk Management of Complex Technical Systems, JSC NIIAS, Moscow, Russian Federation, e-mail: obpronevich@gmail.com

Anna P. Klokova, Postgraduate Student, Russian University of Transport RUT (MIIT), Specialist, Division for Risk Management of Complex Technical Systems, JSC NIIAS, Moscow, Russian Federation, e-mail: i@aklokova.ru

Вклад авторов в статью

Проневич О.Б. Постановка задачи, проведен обзор и анализ математического *UMAP*, определены ключевые параметры *UMAP*, оказывающие влияние на результаты снижения размерности.

Клокова А.П. Проведен обзор методов снижения размерности и обоснован выбор *UMAP* как основного метода снижения размерности, разработан алгоритм отбора признаков и проведения нормировки до применения *UMAP*, разработан пример применения *UMAP* для снижения размерности данных, характеризующих состояние электровозов.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов