Гайнетдинова, А. А. Сравнение методов отбора значимых признаков для классификации геомагнитных данных / А. А. Гайнетдинова, А.В. Воробьев // Прикладная математика и вопросы управления. – 2023. – № 4. – С. 46–54. DOI 10.15593/2499-9873/2023.4.02

Библиографическое описание согласно ГОСТ Р 7.0.100–2018
Гайнетдинова, А. А. Сравнение методов отбора значимых признаков для классификации геомагнитных данных / А. А. Гайнетдинова, А. В. Воробьев. – Текст : непосредственный. – DOI 10.15593/2499-9873/2023.4.02 // Прикладная математика и вопросы управления / Applied Mathematics and Control Sciences. – 2023. – № 4. – С. 46–54.

пермский политех

ПРИКЛАДНАЯ МАТЕМАТИКА И ВОПРОСЫ УПРАВЛЕНИЯ
№ 4, 2023
https://ered.pstu.ru/index.php/amcs

Научная статья
DOI: 10.15593/2499-9873/2023.4.02
УДК 519.816

Сравнение методов отбора значимых признаков для классификации геомагнитных данных

А.А. Гайнетдинова, А.В. Воробьев

Уфимский университет науки и технологий, Уфа, Российская Федерация

О СТАТЬЕ

Получена: 09 июля 2023
Одобрена: 11 декабря 2023
Принята к публикации:
14 декабря 2023

Финансирование
Исследование не имело спонсорской поддержки.
Конфликт интересов
Авторы заявляют об отсутствии конфликта интересов.
Вклад авторов
равноценен

Ключевые слова:
геомагнитные данные, полярные сияния, машинное обучение, обработка данных, значимые признаки, метод главных компонент, метод рекурсивного исключения признаков, деревья решений, факторные нагрузки, метод опорных векторов.

АННОТАЦИЯ

Рассматриваются основные этапы обработки и методы отбора признаков для их дальнейшего использования в алгоритмах машинного обучения для построения моделей, которые предназначены для прогнозирования полярных сияний. Целью работы является сравнение методов отбора признаков при построении модели диагностики наличия полярных сияний на основе интеллектуального анализа геомагнитных данных.

В качестве исходных данных для настоящей работы использовались данные обсерватории «Ловозеро» (LOZ) за девять лет (2012–2020 гг.). Отличительной особенностью данных является их разнородность: в наборе содержатся как категориальные (часть которых являются бинарными, а часть – небинарными), так и количественные.

Рассмотрены такие способы отбора признаков, как анализ главных компонент, метод опорных векторов, рекурсивное исключение признаков, алгоритм Extra-Trees. Результаты исследования показали, что использование отобранных признаков на основе анализа в проекции главных компонент позволит преодолеть «проклятье размерности», устранить «шумы» и снизить переобучение модели.

© **Гайнетдинова Алия Айдаровна** – кандидат физико-математических наук, доцент кафедры высокопроизводительных вычислительных технологий и систем, e-mail: gajnetdinova.aa@ugatu.su, ORCID 0000-0002-4424-6104.
Воробьев Андрей Владимирович – доктор технических наук, доцент, профессор кафедры геоинформационных систем, e-mail: geomagnet@list.ru, ORCID 0000-0002-9680-5609.

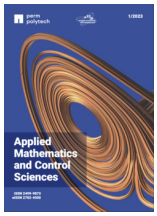
Perm Polytech Style: Gainetdinova A.A., Vorobev A.V. Comparison of features elimination methods for geomagnetic data classification. *Applied Mathematics and Control Sciences.* 2023, no. 4, pp. 46–54. DOI: 10.15593/2499-9873/2023.4.02

MDPI and ACS Style: Gainetdinova, A.A.; Vorobev, A.V. Comparison of features elimination methods for geomagnetic data classification. *Appl. Math. Control Sci.* **2023**, 4, 46–54. https://doi.org/10.15593/2499-9873/2023.4.02

Chicago/Turabian Style: Gainetdinova Aliya A., and Andrei V. Vorobev. 2023. "Comparison of features elimination methods for geomagnetic data classification". *Appl. Math. Control Sci.* no. 4: 46–54. https://doi.org/10.15593/2499-9873/2023.4.02

perm polytech

APPLIED MATHEMATICS AND CONTROL SCIENCES
№ 4, 2023
https://ered.pstu.ru/index.php/amcs



Article
DOI: 10.15593/2499-9873/2023.4.02
UDC 519.816

Comparison of Features Elimination Methods for Geomagnetic Data Classification

A.A. Gainetdinova, A.V. Vorobev

Ufa University of Science and Technology, Ufa, Russian Federation

ARTICLE INFO

Received: 09 July 2023
Approved: 11 December 2023
Accepted for publication:
14 December 2023

Funding
This research received no external funding.
Conflicts of Interest
The authors declare no conflict of interest.
Authors Contributions
equivalent.

Keywords:
geomagnetic data, auroras, machine learning, data processing, significant features, principal component analysis, recursive feature elimination, decision trees, factor loadings, support vector machine.

ABSTRACT

The main stages of processing and feature selection methods for their further use in machine learning algorithms for building models that are designed to predict auroras are considered. The aim of this work is to compare the methods of feature selection when constructing a model for diagnosing the presence of auroras based on the intellectual analysis of geomagnetic data.

Data from the Lovozero Observatory (LOZ) for nine years (2012–2020) were used as data for processing. A distinctive feature of the data is their heterogeneity: the set contains both categorical (binary and non-binary) and quantitative data.

We consider such feature selection methods as principal component analysis, support vector machines, recursive feature elimination, and the Extra-Trees algorithm.

The results of the study showed that the use of selected features based on the analysis in the projection of the principal components will overcome the curse of dimensionality, eliminate noise and reduce model overfitting.

© **Aliya A. Gainetdinova** – CSs of Physics and Mathematics Sciences, Associate Professor, Department of High Performance Computing Technologies and Systems, e-mail: gajnetdinova.aa@ugatu.su, ORCID: 0000-0002-4424-6104.
Andrey V. Vorobev – Doctor of Technical Sciences, Docent, Professor of the Geoinformation Systems Department, e-mail: geomagnet@list.ru, ORCID: 0000-0002-9680-5609.

Введение

Обеспечение эффективного прогнозирования и диагностирования полярных сияний является актуальной задачей, так как полноценный мониторинг космической погоды позволяет предотвратить возможные негативные последствия [1–3].

Применение методов интеллектуального анализа данных и моделей машинного обучения в задачах диагностики полярных сияний требует отбора значимых признаков геомагнитных данных в целях устранения мультиколлинеарности переменных модели [4; 5]. Данное свойство негативно влияет на качество обучения модели, так как вносит дополнительные «шумы» и приводит к неустойчивости прогностической модели [6].

Специфика описания геомагнитных данных предполагает использование ряда показателей (подробнее – в следующем разделе), как категориальных (часть которых являются бинарными, а часть – небинарными), так и количественных. Такой набор данных предполагает комбинирование методов отбора признаков, что обусловливает актуальность исследований в данном направлении.

Целью работы является сравнение методов отбора признаков при построении модели диагностики наличия полярных сияний на основе интеллектуального анализа геомагнитных данных. Для проведения расчетов использовались библиотеки языка Python (pandas, scikit-learn).

Описание исходных данных

В качестве исходных данных для настоящей работы использовались данные обсерватории «Ловозеро» (LOZ) за девять лет (2012–2020 гг.). Результаты синхронных наблюдений полярных сияний и вариаций геомагнитного поля (ГМП) в окрестности LOZ представляются в виде аскаплотов (рис. 1), публикуемых на сайте Полярного геофизического института [7].

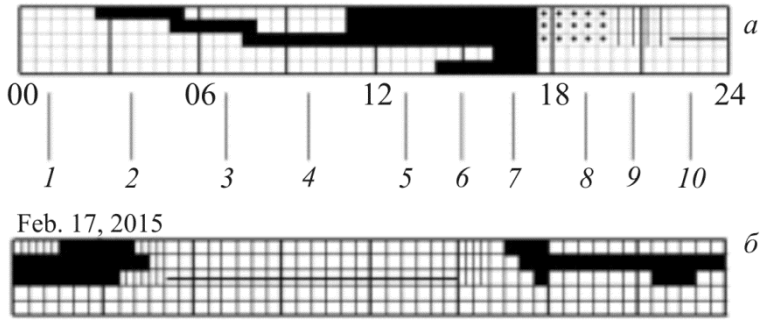


Рис. 1. Формат представления данных в виде аскаплота: *1* – сияние не наблюдается; *2* – сияние в северной области; *3* – сияние в зените; *4* – сияние на юге; *5* – сияние в зените, северной и южной областях; *6* – умеренное сияние в зените, кроме этого, свечение присутствует в северной и южной областях; *7* – сильное сияние в зените, кроме этого, свечение присутствует в северной и южной областях; *8* – частичная облачность; *9* – сплошная облачность; *10* – регистрация не проводилась (*а*); пример аскаплота обсерватории LOZ за 17.02.2015 г., источник PGI Geophysical data [7] (*б*)

Такое представление требует оцифровки, результатом которой являются электронные таблицы вида, соответствующего табл. 1.

Данные о геомагнитных вариациях (ГМВ) в месте наблюдения сияний (с. Ловозеро) доступны на сайте проекта SuperMag [8], который помимо сбора и хранения геомагнитных дан-

ных реализует также некоторые процедуры их предварительной обработки, например, исключение суточной составляющей вариаций ГМП, годового тренда и константы смещения [9].

Таблица 1

Фрагмент оцифрованного аскаплота за 17.02.2015
(см. рис. 1, *б*) *(авторские результаты)*

Дата	UTC	Область проявления сияний относительно места наблюдений				
		север	зенит	юг	сияние в зените, а также на севере и юге	
					зенит (умеренное)	зенит (сильное)
17.02.2015	00:00	×	1	×	0	0
17.02.2015	00:30	×	1	×	0	0
17.02.2015	01:00	×	1	×	0	0
…	...	…	…	…	…	…
17.02.2015	23:00	0	1	0	0	0
17.02.2015	23:30	0	1	0	0	0

Примечание: 1 – сияние наблюдалось; 0 – сияние не наблюдалось; × – полная или частичная облачность.

Приведем подробное описание используемых в дальнейшей работе данных.

DATETIME – признак, представляющий дату и время наблюдения или измерения, полезен для анализа временной зависимости данных или для организации данных в хронологическом порядке.

SME и SMR – это признаки, рассчитываемые в рамках проекта SuperMAG, связаны с геомагнитной активностью в авроральной и экваториальных зонах соответственно.

AP – это признак, связанный интегральной мощностью светимости в авроральной зоне.

N_{LOZ}, E_{LOZ}, Z_{LOZ}, H_{LOZ}, F_{LOZ}, D_{LOZ}, I_{LOZ} – компоненты магнитного поля, измеряемого на обсерватории LOZ. N_{LOZ} представляет северную (горизонтальную) компоненту, E_{LOZ} – восточную (горизонтальную) компоненту, Z_{LOZ} – вертикальную компоненту, H_{LOZ} – горизонтальную полную интенсивность, F_{LOZ} – наклонную интенсивность и D_{LOZ} – деклинацию, I_{LOZ} – инклинацию.

ΔN_{LOZ}, ΔE_{LOZ}, ΔZ_{LOZ} – изменение компонент магнитного поля LOZ; ΔN_{LOZ} – изменение северной (горизонтальной) компоненты, ΔE_{LOZ} – изменение восточной (горизонтальной) компоненты и ΔZ_{LOZ} – изменение вертикальной компоненты.

$|\Delta N_{LOZ}|$, $|\Delta E_{LOZ}|$, $|\Delta Z_{LOZ}|$ – абсолютное значение изменения компонент магнитного поля LOZ. Они могут быть полезны для оценки амплитуды изменений в магнитном поле.

$|d\Delta Z_{LOZ}/dt|$, $|d\Delta N_{LOZ}/dt|$, $|d\Delta E_{LOZ}/dt|$ – эти признаки отображают разницу между соседними значениями изменения компонент магнитного поля LOZ. Они могут быть использованы для анализа изменчивости или трендов в изменении магнитного поля.

dN_{LOZ}/dt, dE_{LOZ}/dt, dZ_{LOZ}/dt, dH_{LOZ}/dt, dF_{LOZ}/dt – это признаки, отражающие изменение скорости в соответствующих компонентах магнитного поля.

Обзор методов машинного обучения для отбора признаков

Для выделения признаков, имеющих наибольшее влияние, могут использоваться методы фильтрации (коэффициент корреляции, абсолютное отклонение и др.), методы обертки (прямой отбор признаков, последовательный отбор признаков и др.), а также встроенные методы (регуляризация Lasso, метод с использованием случайного леса и др.) [5].

Для данного исследования были выбраны следующие методы машинного обучения.

Метод главных компонент (principal components analysis – PCA) впервые был предложен в работе К. Пирсона [10]. PCA преследует несколько целей:

- снижение размерности данных;
- построение ординации объектов;
- исследование связей между переменными.

В основе метода лежит спектральное разложение ковариационной матрицы (см., например, [11]). Для этого предварительно данные нормируются и центрируются, после чего находится матрица ковариаций и строится ее SVD-разложение. Значения полученных сингулярных чисел показывают, каков «разброс» для каждой из компонент, а значит, какая доля полезной информации хранится в ней: чем больше значение, тем выше «вклад» соответствующей компоненты в «разброс» признаков.

Метод опорных векторов (support vector machines – SVM, см. [12]) – это метод обучения с учителем, используемый для классификации и регрессии. Он основывается на поиске оптимальной разделяющей гиперплоскости между двумя классами данных, с максимальным зазором между классами. SVM относится к методам машинного обучения с ядром (kernel), которое позволяет проецировать данные в пространство большей размерности, чтобы найти нелинейную разделяющую гиперплоскость.

Метод рекурсивного исключения признаков (Recursive Feature Elimination – RFE, см. [13]) реализует следующий алгоритм: модель обучается на исходном наборе признаков и оценивает их значимость, затем исключается один или несколько наименее значимых признаков, модель обучается на оставшихся признаках, и так далее, пока не останется заданное количество лучших признаков.

Ансамблевые алгоритмы на основе деревьев решений, такие как случайный лес (Random Forest), также позволяют оценить важность признаков. В настоящей работе для сравнения работы алгоритмов был использован алгоритм Extra-Trees (Extremely Randomized Trees, см. [14]). Данный алгоритм строит ансамбль необрезанных деревьев решений или регрессии в соответствии с классической нисходящей процедурой. Его два основных отличия от других ансамблевых алгоритмов на основе деревьев заключаются в том, что он разделяет узлы, выбирая точки среза полностью случайным образом, и что он использует всю обучающую выборку (а не реплику начальной загрузки) для выращивания деревьев.

Предварительный анализ данных

Имеющие место пропуски в данных, связанные с кратковременными отказами магнитометра LOZ, здесь восстанавливались линейной интерполяцией без видимого ущерба точности информационного сигнала. Информация, утраченная по причине более длительных эпизодов неработоспособного состояния системы, исключалась из генеральной совокупности и не рассматривалась. Исключались также значения, имеющие выраженный аномальный характер на фоне соответствующих им выборок.

Перед началом работы все данные было необходимо преобразовать в количественные и стандартизовать.

Категориальные признаки, принимающие два значения (т.е. бинарные признаки) и большее количество значений, обрабатываются по-разному. Значения бинарных признаков просто заменяют на 0 и 1. К небинарным признакам применяют метод векторизации [15].

Для нормализации данных использовалось масштабирование функций с помощью MinMaxScaler из пакета sklearn.preprocessing (см. [16]): эта функция масштабирует данные таким образом, чтобы они находились между нулем и единицей.

С использованием метода главных компонент были определены корреляционные связи в данных. Для исследуемого набора использованы две главные компоненты, дающие соответственно 47,08 и 33,96 % вклада в разброс признаков. На диаграмме рассеяния (рис. 2) по оси x отложена первая главная компонента, по оси y – вторая.

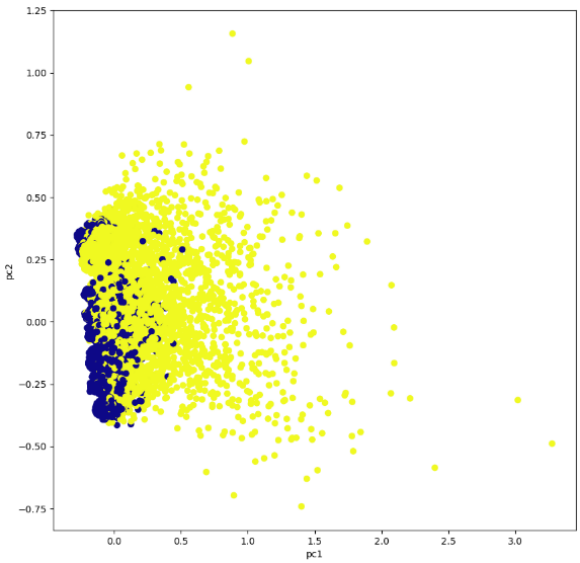


Рис. 2. Двумерная диаграмма рассеяния для исследуемого набора данных *(авторские результаты)*

В рассматриваемом двумерном пространстве эти два класса разделены нечетко. Можно предположить, что простых линейных классификаторов будет недостаточно для решения задачи классификации.

Полученные результаты

Для рассматриваемого набора данных результат метода PCA таков: 5 первых компонент дают 89,45 % вклада, причем первая компонента дает 57,25 %. Полученные главные компоненты можно трактовать как сложные признаки, и для их интерпретации используются факторные нагрузки: признаки изменяются вдоль компоненты тем сильнее, чем больше модуль их факторной нагрузки; знак факторной нагрузки означает направление изменения исходного признака вдоль главной компоненты. Применительно к исследуемым данным PCA позволяет выделить следующие значимые признаки: $d\Delta N_{LOZ}/dt$, $d\Delta Z_{LOZ}/dt$, $d\Delta F_{LOZ}/dt$, SME, AP.

Результат применения метода опорных векторов – выделение признаков AP, SME, E_{LOZ}, D_{LOZ}, H_{LOZ}.

Метод RFE в комбинации с логистической регрессией выделяет признаки dH_{LOZ}/dt, dN_{LOZ}/dt, ΔZ_{LOZ}, F_{LOZ}, dF_{LOZ}/dt.

Признаки, выделенные алгоритмом Extra Trees Classifier, – ΔZ_{LOZ}, dE_{LOZ}/dt, dH_{LOZ}/dt, Z_{LOZ}, LOZ_Z, dF_{LOZ}/dt.

Для оценки качества моделей были вычислены коэффициенты детерминации R^2 [17; 18] (табл. 2): значение, близкое к единице, указывает на то, что метод лучше соответствует данным, тогда как значение, близкое к нулю, означает, что метод не объясняет изменчивость данных.

Таблица 2

Коэффициенты детерминации методов *(авторские результаты)*

Метод	Признак			
	RFE	Extra-Trees	PCA	SVM
R^2	0,43	0,57	0,89	0,49

Исследование, связанное с полученными результатами, приведено в [19]: на основе выбранных признаков с применением байесовского вывода предложен подход к диагностике полярных сияний, точность которого составила не менее 86 % при использовании нескольких локальных предикторов и ~80 % при использовании нескольких глобальных индексов геомагнитной активности.

Заключение

В этом исследовании проведено эмпирическое сравнение четырех методов выбора признаков. Гипотеза исследования предполагала различия в эффективности рассматриваемых методов выбора признаков. Результаты показали, что использование отобранных признаков на основе анализа в проекции главных компонент позволит преодолеть «проклятье» размерности, устранить «шумы» и снизить переобучение модели.

Кроме того, можно предположить, что целесообразно составить какие-либо комбинации алгоритмов машинного обучения с целью получения более эффективных методов отбора признаков (например, как в [20]) или построить более сложные и сильные синтетические предикторы, на существование которых однозначно указывают результаты анализа главных компонент.

Список литературы

1. Пилипенко В.А. Воздействие космической погоды на наземные технологические системы // Солнечно-земная физика. – 2021. – Т. 7, № 3. – С. 73–110. DOI: 10.12737/szf73202106

2. Влияние космической погоды на надежность функционирования железнодорожного транспорта в арктической зоне России / И.Н. Розенберг, А.Д. Гвишиани, А.А. Соловьев, В.А. Воронин, В.А. Пилипенко // Железнодорожный транспорт. – 2021. – № 12. – С. 48–54.

3. Демьянов В.В., Ясюкевич Ю.В. Космическая погода: факторы риска для глобальных навигационных спутниковых систем // Солнечно-земная физика. – 2021. – Т. 7, № 2. – С. 30–52. DOI: 10.12737/szf72202104

4. Храмов А.Г. Методы и алгоритмы интеллектуального анализа данных. – Самара: Изд-во Самарского университета, 2019. – 176 с.

5. Zheng A., Casari A. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. – O'Reilly Media, Inc., 2018. – 218 p.

6. Mitigating the multicollinearity problem and its machine learning approach: a review / J.Y.-L. Chan, S.M.H. Leow, K.T. Bea, W.K. Cheng, S.W. Phoong, Z.-W. Hong, Y.-L. Chen // Mathematics. – 2022. – Vol. 10, iss. 8. – No. 1283. DOI: 10.3390/math10081283

7. Archive of PGI Geophysical Data [Электронный ресурс] / Федеральное государственное бюджетное научное учреждение «Полярный геофизический институт». – URL: http://pgia.ru/lang/ru/archive_pgi (дата обращения: 08.07.2023).

8. Сервис SuperMAG [Электронный ресурс] / John Hopkins Applied Physics Laboratory. – URL: https://supermag.jhuapl.edu/mag (дата обращения: 08.07.2023).

9. Gjerloev J.W. The SuperMAG data processing technique // Journal of Geophysical Research: Space Physics. – 2012. – Vol. 117, iss. A9. – P. A09213. DOI: 10.1029/2012JA017683

10. Pearson K. On lines and planes of closest fit to systems of points in space // Philosophical Magazine. – 1901. – Vol. 2. – P. 559–572.

11. Прикладная статистика. Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1989. – 607 с.

12. Nefedov A. Support Vector Machines: A Simple Tutorial [Электронный ресурс]. – 2016. – URL: https://svmtutorial.online/SVM_tutorial.pdf (дата обращения: 08.07.2023).

13. Gene Selection for Cancer Classification Using Support Vector Machines / I. Guyon, J. Weston, S. Barnhill, V. Vapnik // Machine Learning. – 2002. – Vol. 46 (1). – P. 389–422. DOI: 10.1023/A:1012487302797

14. Geurts P., Ernst D., Wehenkel L. Extremely randomized trees // Machine Learning. – 2006. – Vol. 63. – P. 3–42. DOI: 10.1007/s10994-006-6226-1

15. Moffitt C. Guide to encoding categorical values in Python [Электронный ресурс] / Practical Business Python. – URL: https://pbpython.com/categorical-encoding.html (дата обращения: 08.07.2023).

16. Предварительная обработка данных [Электронный ресурс] scikit-learn developers. – URL: https://scikit-learn.ru/6-3-preprocessing-data/ (дата обращения: 08.07.2023).

17. Бахрушин В.Е. Методы оценивания характеристик нелинейных статистических связей // Системные технологии. – 2011. – № 2 (73). – С. 9–14.

18. Chicco D., Warrens M.J., Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation // PeerJ Computer Science. – 2021. – Vol. 7. – P. e623. DOI: 10.7717/peerj-cs.623

19. Локальная диагностика наличия полярных сияний на основе интеллектуального анализа геомагнитных данных / А.В. Воробьев, А.А. Соловьев, В.А. Пилипенко, Г.Р. Воробьева, А.А. Гайнетдинова, А.Н. Лапин, В.Б. Белаховский, А.В. Ролдугин // Солнечно-земная физика. – 2023. – Т. 9, № 2. – С. 26–34. DOI: 10.12737/szf-92202303

20. Комбинированная схема отбора признаков для разработки банковских моделей / С.В. Афанасьев, Д.М. Котерева, А.А. Мироненков, А.А. Смирнова // Финансы: теория и практика. – 2023. – Т. 27, № 1. – С. 103–115. DOI: 10.26794/2587-5671-2023-27-1-103-115

References

1. Pilipenko V.A. Space weather impact on ground-based technological systems. *Solar-Terrestrial Physics*, 2021, vol. 7, no. 3, pp. 68-104. DOI: 10.12737/szf73202106

2. Rozenberg I N., Gvishiani A. D., Solov'ev A. A., Voronin V. A., Pilipenko V. A. Vliianie kosmicheskoi pogody na nadezhnost' funktsionirovaniia zheleznodorozhnogo transporta v arkticheskoi zone Rossii. *Zheleznodorozhnyi transport*, 2021, № 12, pp. 48–54.

3. Demyanov V.V., Yasyukevich Yu.V. Space weather: risk factors for global navigation satellite systems. *Solar-Terrestrial Physics*, 2021, vol. 7, no. 2, pp. 28-47.

4. Khramov A. G. Metody i algoritmy intellektual'nogo analiza dannykh. Samara, Samarskii universitet, 2019, 176 p.

5. Zheng A., Casari A. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media, Inc., 2018, 218 p.

6. Chan J.Y.-L.; Leow S.M.H.; Bea K.T.; Cheng W.K.; Phoong S.W.; Hong Z.-W.; Chen Y.-L. Mitigating the multicollinearity problem and its machine learning approach: a review. *Mathematics,* 2022, vol. 10, no. 8, p. 1283. DOI: 10.3390/math10081283

7. Archive of PGI Geophysical Data, available at: http://pgia.ru/lang/ru/archive_pgi (Accessed 08 July 2023)

8. SuperMAG, available at: https://supermag.jhuapl.edu/mag (Accessed 08 July 2023)

9. Gjerloev J. W. The SuperMAG data processing technique. *Journal of Geophysical Research: Space Physics,* 2012, vol. 117, no. A9, p. A09213. DOI: 10.1029/2012JA017683

10. Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 1901, vol. 2, pp. 559–572.

11. Aivazian S. A., Bukhshtaber V. M., Eniukov I. S., Meshalkin L. D. Prikladnaia statistika. Klassifikatsiia i snizhenie razmernosti. Moscow, Finansy i statistika, 1989, 607 p.

12. Nefedov, A. Support Vector Machines: A Simple Tutorial, 2016. Available at: https://svmtutorial.online/SVM_tutorial.pdf (Accessed 08 July 2023)

13. Guyon I., Weston J., Barnhill S., Vapnik V. Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning,* 2002, vol. 46 (1), pp. 389–422. DOI: 10.1023/A:1012487302797

14. Geurts P., Ernst D., Wehenkel L. Extremely randomized trees. *Machine Learning*, 2006, vol. 63, pp. 3–42. DOI: 10.1007/s10994-006-6226-1

15. Moffitt, C. Guide to encoding categorical values in Python. Available at: https://pbpython.com/categorical-encoding.html (Accessed 08 July 2023)

16. Scikit-learn: 6.3. Preprocessing data. Available at: https://scikit-learn.ru/6-3-preprocessing-data/ (Accessed 08 July 2023)

17. Bakhrushin V. E. Metody otsenivaniia kharakteristik nelineinykh statisticheskikh sviazei [Methods for estimating characteristics of nonlinear statistical couplings]. System technologies, 2011, no. 2 (73), pp. 9–14.

18. Chicco D., Warrens M. J., Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 2021, vol. 7, p. e623. DOI: 10.7717/peerj-cs.623

19. Vorobev A.V., Soloviev A.A., Pilipenko V.A., Vorobeva G.R., Gainetdinova A.A., Lapin A.N., Belakhovsky V.B., Roldugin A.V. Local diagnostics of aurora presence based on intelligent analysis of geomagnetic data. *Solar-Terrestrial Physics*, 2023, vol. 9, no. 2, pp. 22-30. DOI: 10.12737/szf-92202303.

20. Afanasyev S.V., Kotereva D. M., Mironenkov A.A., Smirnova A.A. Combined Feature selection scheme for banking modeling. *Finance: Theory and Practice*, 2023, vol. 27, no. 1, pp. 103–115. DOI: 10.26794/2587-5671-2023-27-1-103-115.