УДК 621/37 DOI: 10.24412/2304-0297-2023-1-275-283

Выбор ансамблевых моделей машинного обучения для прогнозирования полосы когерентности трансионосферных каналов связи

Е.М. Антропова, Н.А. Конкин

Поволжский государственный технологический университет 424000, г. Йошкар-Ола, пл. Ленина д.3.
E-mail: konkinna@volgatech.net

Представлен анализ ансамблевых моделей машинного обучения с целью прогнозирования значений полосы когерентности трансионосферных каналов связи. Получены результаты прогнозирования значений полосы когерентности с помощью ансамблевых моделей машинного обучения XGBoost, AdaBoost и GBR. Разработана программа на языке программирования Python подбора моделей машинного обучения для решения задачи прогнозирования значений полосы когерентности.
Ключевые слова: машинное обучение, модели машинного обучения, полоса когерентности, прогнозирование полосы когерентности, Python.

Selection of ensemble machine learning models for predicting the coherence band of trans-ionospheric communication channels

E.M Antropova., N.A .Konkin

Volga State Technological University

The analysis of ensemble models of machine learning is presented in order to predict the values of the coherence band of trans-ionospheric communication channels. The results of predicting the values of the coherence band using the ensemble machine learning models XGBoost, AdaBoost and GBR are obtained. A program has been developed in the Python programming language for selecting machine learning models to solve the problem of predicting the values of the coherence band.
Keywords: machine learning, machine learning models, coherence band, coherence band prediction, Python.

Введение.

Машинное обучения является совокупностью различных математических, вычислительных и статистических методов, которые лежат в основе алгоритмов, позволяющих решить задачи обработки данных, в частности временных рядов. Одним из преимуществ машинного обучения заключается в том, что обработка временных рядов производится на основе входных данных модели машинного обучения, а не за счет заранее известной математической модели. Данный подход является универсальным с точки зрения обработки разнородных данных. В ходе исследования в первую очередь были изучены научные работы и статьи, в которых рассматривался анализ временных рядов, так как и значения полосы когерентности [7-9] трансионосферных каналов связи являются временными рядами. Было определено, что наиболее популярными являются ансамблевые методы машинного обучения [1-3]. Рассмотрено применение таких моделей, как градиентный спуск, деревья принятий

решений, XGBoost (eXtreme Gradient Boosting), ARMA (AutoRegressive–Moving-Average) и других моделей. Таким образом, на основе таких критериев, как возможность работы с временными (регрессии), высокая точность прогнозирования и быстрая обучаемость в текущей работе, применены ансамблевые модели машинного обучения XGBoost[4], AdaBoost[5] и GB (Gradient boosting)[6] для выявления закономерностей в значениях полосы когерентности спутниковых линий связи и их прогнозирования.

Цели и задачи исследования.

Целью работы является выбор ансамблевых моделей машинного обучения для прогнозирования временных рядов полосы когерентности трансионосферных каналов связи.

Задачи исследования:

1. Разработать методику краткосрочного (в пределах одного дня) прогнозирования значений полосы когерентности на основе нескольких моделей машинного обучения с возможностью последующего сравнительного анализа.
2. Создать программный комплекс реализации методики краткосрочного прогнозирования значений полосы когерентности на языке программирования Python.
3. На основе результатов прогноза определить с помощью метрик средней абсолютной ошибки и коэффициента детерминации более точный алгоритм машинного обучения.

Метод машинного обучения AdaBoost.

Аддитивное моделирование заключается в том, что в цикле каждый раз добавляется по одной базовой модели и обучение модели проходит на всем объеме учебных данных и решает задачи классификации и регрессии. При бустинге обучаются T число алгоритмов, а затем вычисляется взвешенная сумма по формуле:

$$a(x) = \sum_{i=1}^{T} a_i b_i(x) , \tag{1}$$

где a_i – весовой коэффициент, b_i – алгоритм модели машинного обучения.

В случае решений задач регрессии критерий отклонений обычно формируют при использовании квадратичной функций потерь:

$$Q_T = \frac{1}{2}\sum_{i=1}^{i}(y_i - \sum_{j=1}^{T} a_j b_j(x_i))^2 = \frac{1}{2}\sum_{i=1}^{i}(y_i - \sum_{j=1}^{T-1} b_j(x_i) - b_T(x_i))^2 \tag{2}$$

Далее определяется выбор семейства алгоритмов $b_j(x)$ задач регрессии, частным случаем которых являются решающие деревья. При выборе решающего дерева при квадратичной функции потерь в листах сохраняются средние значения целевых меток y_i объектов. Если представить, что в некоторую листовую вершину попали M количество объектов обучающей выборки R_V, то оптимальное значение для описания можно определить как:

$$\begin{gathered} Q = \frac{1}{2}\sum_{ieR_V}(y_i - c)^2 \to \min \\ \frac{dQ}{dc} = \sum_{ieR_V}(y_i - c) = 0 \end{gathered} \tag{3}$$

откуда:

$$c = \frac{1}{|R_V|} \sum_{ieR_V} y_i \tag{4}$$

а R_V определяется как:

$$R_V = \left\{ \left(x_i, y_i \right)_{i=1}^{M} \right\} \tag{5}$$

Метод машинного обучения Gradient boosting.

Принцип работы градиентного бустинга заключается в совокупности слабых моделей прогнозирования, которые представляют собой деревья решений. Далее приведен алгоритм работы градиентного бустинга. Входными данными алгоритма являются обучающая выборка $\{(x_i, y_i,)\}_{i=1}^{n}$, дифференцируемая функция потерь $L(y, F(x))$ и количество итераций M. Инициализируем модель:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma), \tag{6}$$

где L функция потерь и рассчитывается как:

$$L = (y_i - \gamma)^2 \tag{7}$$

В (7) значение γ минимизирует градиент и вычисляется следующим образом:

$$\frac{\partial}{\partial \gamma} \sum_{i=1}^{n} L = \frac{\partial}{\partial \gamma} \sum_{i=1}^{n} (y_i - \gamma)^2 = -2 \frac{\partial}{\partial \gamma} \sum_{i=1}^{n} (y_i - \gamma) = -2 \frac{\partial}{\partial \gamma} \sum_{i=1}^{n} y_i + 2n\gamma \tag{8}$$

Далее начинается цикл для *m=1 до M*, первым идет вычисление остатков:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \quad for\, i = 1, \ldots, n., \tag{9}$$

и обучение на тренировочном наборе $\{(x_i, r_{im})\}_{i=1}^{n}$. Для определения длины шага используется формула:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)), \tag{10}$$

где $h_m(x)$ лучшая функция. В конце цикла модель обновляется:

$$F_m(x) = F_{m-1}(x) + \underset{\gamma}{\gamma_m} h_m(x) \tag{11}$$

На выходе функция будет иметь вид $F_M(x)$.

Метод машинного обучения XGBoost.

В основе модели XGBoost находится алгоритм деревьев принятия решений, принцип модели основан на экстремальном повышении градиента. XGBoost используется для решения различных задач с контролируемыми обучающими данными x_i для прогнозирования целевой переменной y_i. При выполнении обучения модель совершает математические операции, в результате которых выполняется прогнозирование y_i.

Задача обучения модели XGBoost заключается в нахождении наилучшего подхода к обучению целевой функции θ. Для целевых функций характерны две части: потери при обучении (loss function) и регуляция (regularization term):

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \omega(f_i) \quad (12)$$

Где L – функция потерь при обучении, и Ω – параметр регуляции. Потери при обучении определяют на сколько точно прогнозирует модель по отношению к данным. Параметр L оценки потерь может имеет различные представления, например, среднеквадратичной ошибкой (MSE) и задается следующим образом:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (13)$$

Если рассмотрим использование среднеквадратичной ошибки в качестве нашей функции потерь, то целевая функция имеет вид:

$$obj^{(t)} = \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(t-1)} + f_i(x_i)))^2 + \sum_{i=1}^{t} \omega(f_i) = \sum_{i=1}^{n} [2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i)^2] + \omega(f_t) + const\,, \quad (14)$$

где $\omega(f_i)$ сложность модели (функция регуляризации), $f_t(x_i)$ – оценка точности листа *t*-го дерева. Функция регуляризации позволяет избегать переобучения.

Методика эксперимента.

Пошаговый алгоритм эксперимента разделен на 4 основных этапа и представлен на рисунке 1. На этапе 1 данные для расчета полосы когерентности (ПК) получены средствами навигационной системой ГЛОНАСС, средствами сети референсных станций SamrtNet. Навигационные данные аккумулируются на сервере ПГТУ в формате обмена данными спутниковых навигационных приемников – RINEX (Receiver Independent Exchange Format). Данный формат дает возможность производить пост-обработку входных данных для выполнения дальнейших вычислений. На этапе 2 рассчитываются значения полосы когерентности, они сохраняются в локальной базе данных кафедры. Этап 3 включает процессы подготовки датасета значений полосы когерентности и тренировочной, и тестовой выборок для проведения эксперимента. Процесс формирования датасета представляет собой предварительную обработку «сырых» значений ПК с точки зрения интерполяции ошибочно рассчитанных значений или выбросов (аномалий), а также процесс аппроксимации временного ряда. Тренировочные и тестовые выборки необходимы соответственно для тренировки моделей машинного обучения и проверки точности прогноза. На этапе 4 выполняется операции по регулировке временного периода прогноза, обучению и тестированию моделей машинного обучения, сравнительному анализу результатов прогнозирования моделей машинного обучения по метрикам средней абсолютной ошибки и коэффициента детерминации. Средняя абсолютная ошибка – мера ошибки между парными наблюдениями, которая определяется как:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \quad (15)$$

Коэффициент детерминации – определение степени схожести временного шага с моделью путем оценки дисперсии случайной величины и дисперсии ошибки модели:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}\,, \quad (16)$$

где $SS_{res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ – сумма квадратов остатков регрессии, $y_i - \hat{y}_i$ – разность значений реального и спрогнозированного временного хода полосы когерентности. $SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ – общая сумма квадратов (реального временного хода ПК), $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ – среднее значение реального временного ряда.

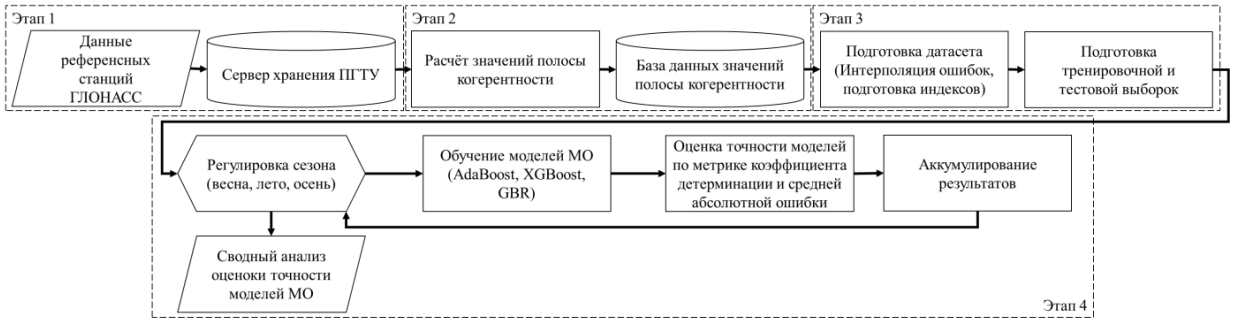

Рис. 1. Алгоритм эксперимента по выбору ансамблевых моделей машинного обучения и прогнозированию значений полосы когерентности трансионосферных каналов связи.

Результаты эксперимента.

Экспериментальные результаты мониторинга параметров трансионосферного радиоканала получены в результате автоматической обработки данных ГЛОНАСС из Банка данных Поволжского государственного технологического университета. Банк данных получен на оборудовании станции Leica GR30 в г. Йошкар-Оле в Поволжском государственном технологическом университете. Статистический анализ и обучение выборки в модели машинного обучения производились на основе 270-дневной базе данных с шагом 30 минут.

Датасет составил 12960 значений. Для повышения точности моделей машинного обучения использованы дополнительные независимые признаки, такие как час, день месяца, квартал, месяц, день года, день, неделя года. На рис. 2-4 представлены результаты прогнозирования в виде зависимости абсолютного значения полосы когерентности от времени моделей XGBoost, AdaBoost и GBR (Gradient Boosting Regression) для трех сезонов года: весны, лета и осени. На рисунках модели отмечены пунктирными линиями, оригинальная линия ряда — сплошной, синим заполнением обозначен доверительный интервал ряда, который построен скользящим окном с шагом в два отсчета.

Шаг окна выбран на основе автокорреляционной функции исследуемого ряда ПК. На рис. 5 представлен график для сравнительного анализа результатов прогнозирования, гистограммы отображают значения коэффициента детерминации, линии – среднюю абсолютную ошибку.

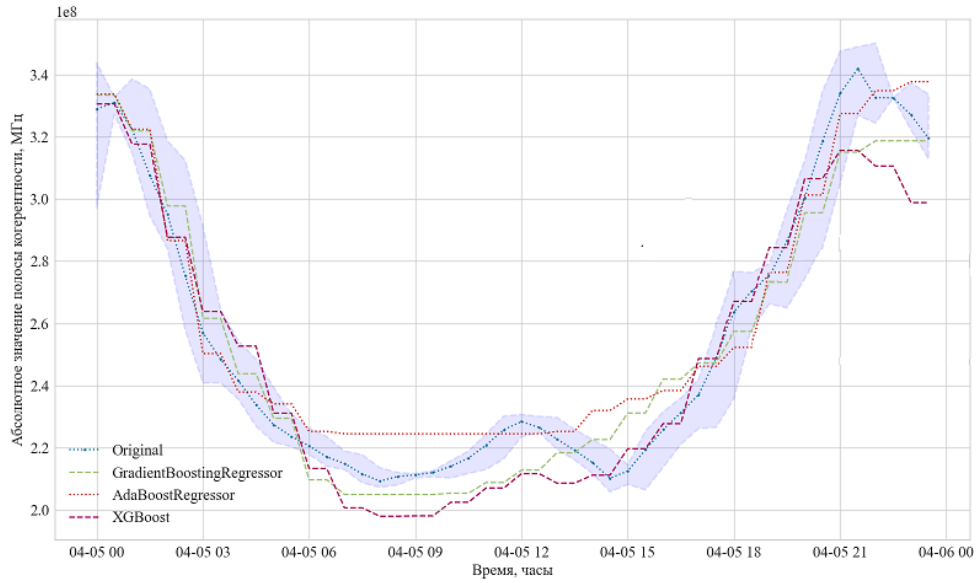

Рис. 2. Результаты прогнозирования значений полосы когерентности с помощью методов машинного обучения XGBoost, AdaBoost и Gradient Boosting трансионофсерных каналов связи в период весны.

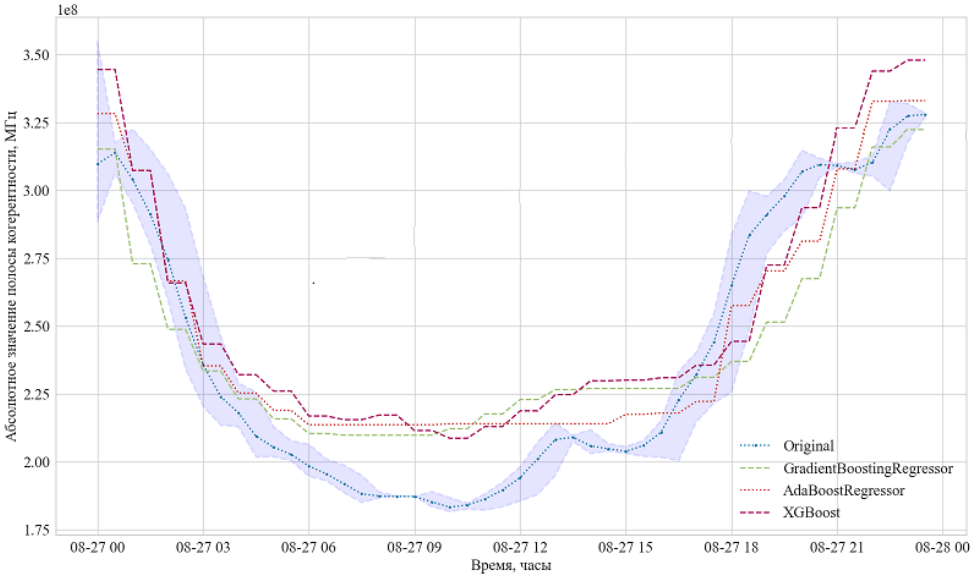

Рис. 3. Результаты прогнозирования значений полосы когерентности с помощью методов машинного обучения XGBoost, AdaBoost и Gradient Boosting трансионофсерных каналов связи в период лета.

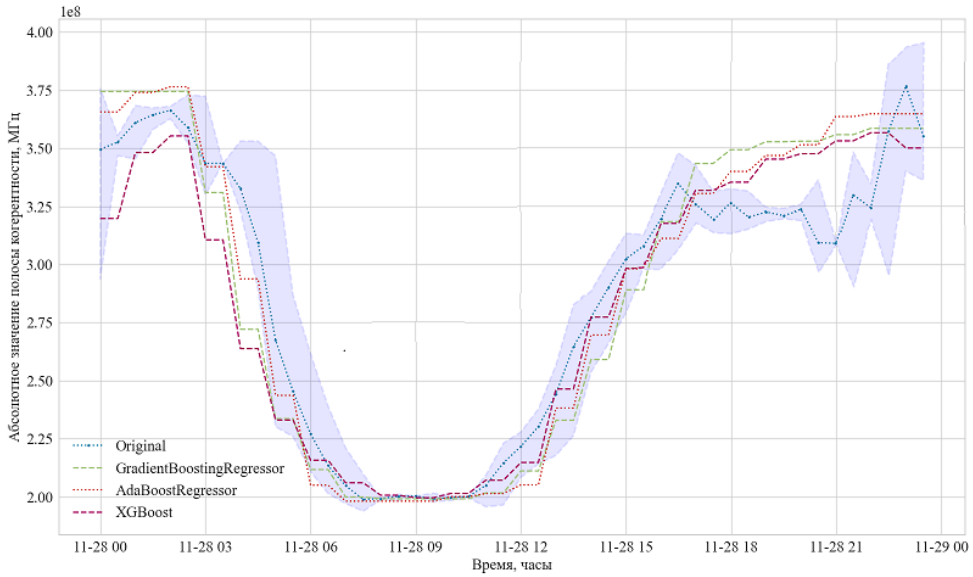

Рис. 4. Результаты прогнозирования значений полосы когерентности с помощью методов машинного обучения XGBoost, AdaBoost и Gradient Boosting трансионофсерных каналов связи в период осени.

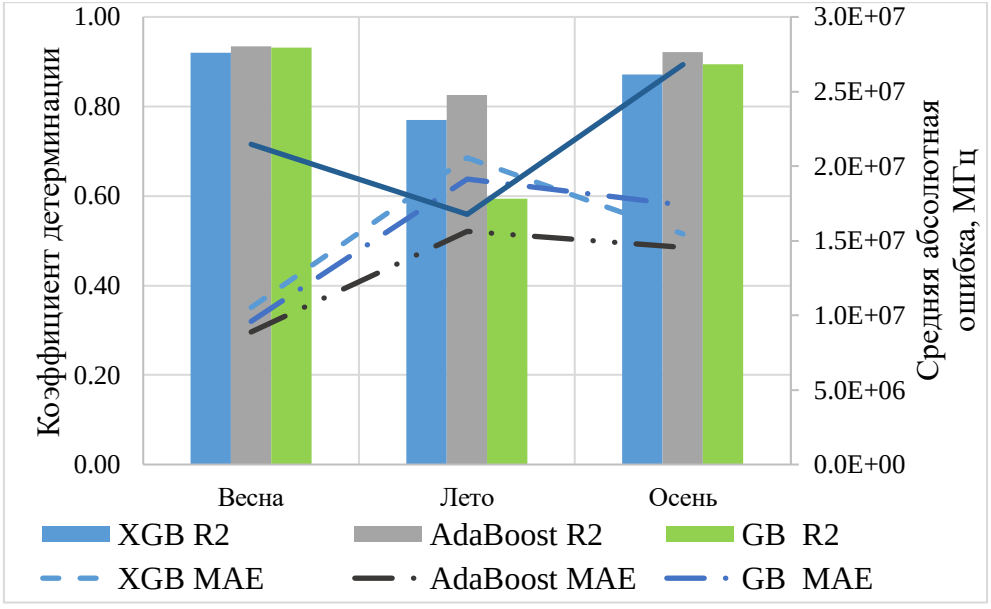

Рис. 5. График оценки точности прогнозов исследуемых моделей машинного обучения по метрикам средней абсолютной ошибки и коэффициента детерминации.

Далее представлены оценки усредненного значения доверительного интервала, средняя абсолютная ошибка и коэффициент детерминации моделей для трех сезонов года.

В весенний сезон доверительный интервал составляет 21 МГц. Средняя абсолютная ошибка для XGBoost составляет 11 МГц, для AdaBoost – 8,9 МГц и для GBR – 9,6 МГц. Коэффициент детерминации для XGBoost составляет 0,92, для AdaBoost – 0,93, для GBR – 0,93.

В летний период доверительный интервал составляет 16 МГц. Средняя абсолютная ошибка для XGBoost составляет 21 МГц, для AdaBoost – 16 МГц, для GBR – 19 МГц.

Коэффициент детерминации для XGBoost составляет 0,77, для AdaBoost – 0,83, для GBR – 0,59.

В осенний сезон доверительный интервал составляет 26 МГц. Средняя абсолютная ошибка для XGBoost составляет 15 МГц, для AdaBoost – 15 МГц, для GBR – 17 МГц. Коэффициент детерминации для XGBoost составляет 0,87, для AdaBoost – 0,92, для GBR – 0,89. Более точной по всем параметрам для всех трех сезонов является AdaBoost.

Заключение.

В ходе эксперимента выполнен выбор ансамблевых моделей машинного обучения для прогнозирования временных рядов значений полосы когерентности трансионосферных каналов связи. Была разработана методика краткосрочного прогнозирования значений полосы когерентности на основе нескольких моделей машинного обучения с возможностью последующего сравнительного анализа. Создан программный комплекс реализации методики краткосрочного прогнозирования значений полосы когерентности на языке программирования Python.

На основе экспериментальных результатов мониторинга параметров трансионосферного радиоканала в различные сезоны года, полученных при автоматической обработке данных ГЛОНАСС из Банка данных ПГТУ, рассчитаны метрики средней абсолютной ошибки и коэффициента детерминации для определения более точного алгоритма машинного обучения. По результатам прогнозирования на основе усредненных оценок точности по метрикам R^2 и *MAE* для весны модель AdaBoost имеет более высокую точность по метрике *MAE* на 1,2 МГц, по R^2 – на 0,01. В летний сезон модель AdaBoost имеет более высокую точность по метрике *MAE* на 4,2 МГц, по R^2 – на 0,14. Осенью модель AdaBoost имеет более высокую точность по метрике *MAE* на 1,9 МГц, по R^2 – на 0,04. Весной и осенью ошибка (*MAE*) прогнозирования для всех моделей укладываются в пределе доверительного интервала, в то время как летом доверительный интервал не превышает только ошибка модель AdaBoost. Данная особенность связана с зашумлённостью исследуемых сезонных временных рядов. При анализе коэффициента детерминации (R^2) учитывалось, что модель со значением коэффициента детерминации выше 0,8 имеет достаточно высокую точность. Этому критерию соответствует только модель AdaBoost со следующими результатами R^2: весной — 0,93, летом — 0,83, осенью — 0,92.

Разработанная методика по выбору моделей машинного изучения даёт возможность прогнозирования полосы когерентности, которая в свою очередь определяет предельную полосу частот, где трансионосферный сигнал спутниковых систем связи имеет минимальные потери и искажения.

Работа выполнена при поддержке гранта № 22-19-00073 Российского научного фонда.

Литература

1. Анализ прогнозирования рядов с помощью автоматизированного машинного обучения в национальной базе данных МКБ-10. 2022. URL: Многошаговое прогнозирования временных рядов с помощью XGBoost. // Towards Data Science URL: https://towardsdatascience.com/multi-step-time-series-forecasting-with-xgboost-65d6820bec39 (дата обращения: 22.10.2022).

2. Многошаговое прогнозирования временных рядов с помощью XGBoost. // Towards Data Science URL: https://towardsdatascience.com/multi-step-time-series-forecasting-with-xgboost-65d6820bec39 (дата обращения: 22.10.2022).

3. Международный журнал прогнозирования. 2022. URL:

https://www.sciencedirect.com/science/article/pii/S0169207021001710 -------> Machine learning algorithms for forecasting and backcasting blood demand data with missing values and outliers: A study of Tema General Hospital of Ghana // ScienceDirect URL: https://www.sciencedirect.com/science/article/pii/S0169207021001710?via%3Dihub (дата обращения: 20.10.2022).

4. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785 (дата обращения 5.02.2023.).

5. Schapire, R. E. (2013). Explaining adaboost. In Empirical inference (pp. 37–52). Springer (дата обращения 2.02.2023.).

6. Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp.1189–1232 (дата обращения 28.02.2023.).

7. Кислицын, А. А. Метод интеллектуального анализа данных для прогнозирования значений полосы когерентности изменяющегося трансионосферного радиоканала / А. А. Кислицын, Н. В. Рябова, Н. А. Конкин // Международная Байкальская молодежная научная школа по фундаментальной физике : Труды XVII Конференции молодых ученых, Иркутск, 05–10 сентября 2022 года. – Иркутск: Федеральное государственное бюджетное учреждение науки Ордена Трудового Красного Знамени Институт солнечно-земной физики Сибирского отделения Российской академии наук, 2022. – С. 361-363. – EDN BJKGKW.

8. Новые возможности систем широкополосной когнитивной связи, работающих в ионосферных КВ-радиоканалах с внутримодовой дисперсией / Д. В. Иванов, В. А. Иванов, Н. В. Рябова, В. В. Овчинников // Радиотехника. – 2022. – Т. 86, № 11. – С. 162-177. – DOI 10.18127/j00338486-202211-23. – EDN JFFPYR.

9. Метод расширения полосы частот систем спутниковой связи путём преодоления дисперсии трансионосферного радиоканала / Д. В. Иванов, В. А. Иванов, Н. В. Рябова, А. А. Кислицын // Радиотехника, электроника и связь: Тезисы докладов VI международной научно-технической конференции, Омск, 06–08 октября 2021 года. – Омск: Омский научно-исследовательский институт приборостроения, 2021. – С. 95-97. – EDN WQJRPD.