Метод анализа данных сложной структуры с элементами машинного обучения

Б.С. Мандрикова [1]

[1] *Институт космофизических исследований и распространения радиоволн ДВО РАН, 684034, Россия, Камчатский край, с. Паратунка, ул. Мирная, д. 7*

Аннотация

Предложен метод анализа данных сложной структуры, основанный на совмещении вейвлет-преобразования и нейронных сетей «Автокодировщик». Метод позволяет изучить структуру данных, выделить аномальные изменения разной формы и длительности и подавить шум. На примере данных сети станций нейтронных мониторов показана эффективность метода. Данные нейтронных мониторов определяют интенсивность вторичных космических лучей и являются одним из ключевых факторов космической погоды. Численная реализация метода позволяет применять его в оперативном режиме, что представляет интерес в задачах анализа природных данных и диагностики катастрофических событий.

Ключевые слова: анализ данных, данные сложной структуры, вейвлет-преобразование, нейронные сети, нейтронные мониторы.

Цитирование: **Мандрикова, Б.С.** Метод анализа данных сложной структуры с элементами машинного обучения / Б.С. Мандрикова // Компьютерная оптика. – 2022. – Т. 46, № 3. – С. 506-512. – DOI: 10.18287/2412-6179-CO-1088.

Citation: Mandrikova BS. A method for analyzing complex structured data with elements of machine learning. Computer Optics 2022; 46(3): 506-512. DOI: 10.18287/2412-6179-CO-1088.

Введение

С ростом технических объектов и развитием научно-технической модернизации в настоящее время все сильнее ощущается проблема создания эффективных методов обработки и анализа сложно структурированных данных. Наиболее сложной задачей является изучение и анализ природных данных ввиду недостаточности знаний об исследуемых процессах и огромного числа воздействующих факторов [1, 2]. Анализ природных данных находит применение в разных сферах человеческой деятельности – физике, биологии, медицине, экономике и др. Особую актуальность имеют методы, направленные на своевременное обнаружение и идентификацию аномалий. Примерами могут служить задачи оперативного распознавания аномалий в данных геофизического мониторинга – предсказание землетрясений, цунами, обнаружение предикторов магнитных бурь, аномалий геологической среды и других катастрофических явлений природы. Необходимость обнаружения аномалий также часто возникает в области медицины, например, для обнаружения и идентификации клинических состояний пациентов. Важным свойством таких методов является их способность к адаптации, обеспечивающая возможность обнаружения и идентификации быстрых изменений состояния системы или объекта, свидетельствующих о возникновении аномалий.

Объектом исследования в работе являются регистрируемые наземными станциями данные интенсивности космических лучей, которые содержат важную информацию о состоянии космической погоды и исследуются огромным числом ученых и научных групп [3 – 5]. Известно, что негативные эффекты космической погоды способны вызвать как техногенные катастрофы (крушение спутников, обесточивание районов, сбои в теле-, радио- и спутниковых системах и т.д.), так и угрозу жизни и здоровью людей (смертельные дозы облучения на бортах самолетов и космических кораблей) [6].

Космические лучи (КЛ) представляют собой потоки высокоскоростных частиц, приходящих из космического пространства на Землю [7]. В работе исследуются вторичные космические лучи – потоки космических частиц, образующиеся в результате их взаимодействия с частицами верхних слоев атмосферы. Вторичные космические лучи регистрируются наземными нейтронными мониторами, число которых в настоящее время превышает 50 станций [8]. Сигнал космических лучей включает периодические (спокойные, не аномальные) вариации и непериодические (аномальные) вариации [7]. К периодическим вариациям относят солнечные циклы, суточные, 27-дневные, 11-летние и т.д. К непериодическим вариациям относят Форбуш-эффекты (ФЭ) и GLE-события (Ground Level Event). ФЭ представляют собой резкие изменения интенсивности потока космических лучей, возникающие во время аномальных процессов в околоземном пространстве, а GLE-событие – резкое сильное протонное возрастание.

Задача обнаружения и распознавания аномалий в космических лучах решается большим числом авторов на протяжении нескольких десятилетий [9 – 12]. Для анализа данных ученые используют физические, математические и статистические модели и методы,

которые становятся все более сложными. Одним из передовых методов анализа данных космических лучей является отечественный метод глобальной съемки, или GSM-метод (Global Survey Method) [9], включающий методы функций связи, траекторных расчетов и сферического анализа. Но сложность расчетов данного метода не позволяет его использовать в оперативном режиме. Авторами [10] предложен алгоритм GLE Alert, способный обнаруживать угрозы космической погоды в режиме реального времени. Данный алгоритм основан на применении пороговых функций и использует скользящее временное окно, что позволяет исследовать динамику процесса и выделить аномальные изменения. Однако дальнейшие разработки [11] показали, что алгоритм не является достаточно эффективным и может давать недостоверный результат – применение алгоритма за 3 года не позволило идентифицировать более 50 % солнечных протонных событий. Другими учеными [12] предлагается применение методологии приближенных байесовских вычислений для оценки параметров модели Форбуш-понижения интенсивности галактических космических лучей. Но данный подход не получил четкой формализации ввиду отсутствия длинных рядов данных, охватывающих длительность Форбуш-понижения, а также отсутствия данных нескольких детекторов с различимой жесткостью откликов [12].

Принимая во внимание указанные проблемы и учитывая сложную нестационарную структуру данных космических лучей, автором предлагается комплексный метод, основанный на применении вейвлет-преобразования и нейронных сетей глубокого обучения. Вейвлет-преобразование позволяет хорошо исследовать детали сигнала по времени и по частоте, извлечь полезную информацию и снизить уровень шума [13, 14]. В работе используется непрерывное вейвлет-преобразование (CWT). Нейронные сети позволяют адаптироваться под изменчивую структуру сложных данных, выделить информативные детали, подавить шум, а также получить результат в оперативном режиме [15 – 17]. Использование нейронной сети «Автокодировщик» дает возможность извлечения зависимостей в данных (за счет минимизации ошибки восстановления) и подавления шума [18]. Данная работа является продолжением исследований [19 – 21]. В статье для повышения эффективности метода, представленного в работе [20], предложена оптимизация нейронной сети на основе регуляризаторов. Используя апостериорный риск, предложен способ оценки порогов, определяющих наличие аномалий в данных. Для снижения вероятности наступления ложной тревоги введено правило, использующее наборы данных и определяющее состояние по совокупности. Для оценки эффективности метода применялись данные ресурса [8].

1. Описание метода

1.1. Аппроксимация данных на основе нейронной сети «Автокодировщик»

Пусть имеем дискретные данные $F[n]$ ($n \in \mathrm{N}$), загрязненные шумом:

$$F[n] = f[n] + V[n],$$

где $f[n]$ – полезный сигнал, $V[n]$ – шум.

Следуя Вальду [22], данные f будем рассматривать как элементы специального множества Θ, без учета распределения вероятности на нем. Тогда, следуя минимаксному критерию [22], задача оценки f состоит в определении оператора решения D, минимизирующего риск

$$r_o(\Theta) = \inf_{D \in O} \sup_{f \in \Theta} E\left\{\left\|\hat{f} - f\right\|^2\right\},$$

где E – математическое ожидание, O – множество операторов, $\hat{f}$ – оценка f.

Рассмотрим в качестве оператора решения D нейронную сеть (НС) «Автокодировщик». Автокодировщик состоит из энкодера и декодера [18].

Энкодер отображает входной вектор F на вектор z:

$$z = h^{(1)}(\omega^{(1)} F + b^{(1)}), \tag{1}$$

где верхний индекс (1) – номер слоя, $h^{(1)}$– передаточная функция, $\omega^{(1)}$ – весовая матрица, а $b^{(1)}$ – вектор смещения.

Декодер отображает закодированное представление z обратно, в оценку исходного входного вектора:

$$\hat{f} = h^{(2)}(\omega^{(2)} z + b^{(2)}), \tag{2}$$

где верхний индекс (2) – номер слоя, $h^{(2)}$ – передаточная функция, $\omega^{(2)}$ – весовая матрица, а $b^{(2)}$ – вектор смещения.

Таким образом, из (1), (2) на основе сети получаем оценку

$$\hat{f} = h^{(2)}(\omega^{(2)}\left(h^{(1)}(\omega^{(1)} F + b^{(1)})\right) + b^{(2)}), \tag{3}$$

где $h^{(1)}$ – передаточная функция энкодера, $\omega^{(1)}$, $\omega^{(2)}$ – матрицы весов, $h^{(2)}$– передаточная функция декодера, $b^{(1)}, b^{(2)}$ – векторы смещения.

Риск оценки f есть

$$r\left(D, f\right) = E\left\{\left\|\hat{f} - f\right\|^2\right\}.$$

Для минимизации риска r может быть выполнена оптимизация сети путем применения регуляризаторов [18]. В этом случае при обучении сети минимизируется функционал $\Im(\aleph)$ по набору параметров $\aleph = \{\omega^{(1)}, \omega^{(2)}, b^{(1)}, b^{(2)}\}$:

$$\Im(\aleph) = \sum_{F \in D_M} L(F, \hat{f}) \to \min_{\aleph},$$

где D_M – множество примеров, L – функция стоимости.

Регуляризаторы включены в функцию стоимости L и позволяют детально изучить признаковое пространство и адаптировать нейросетевую модель под данные. В случае разреженных регуляризаторов функция стоимости L имеет вид [18]:

$$L(F,\hat{f})=\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(F_m[n]-\hat{f}_m[n]\right)^2+\alpha\operatorname{Re}g_\omega+\beta\operatorname{Re}g_\rho,$$

где M – количество примеров, N – размерность обучающих данных, α и β – коэффициенты регуляризации разреженности и регуляризации весов соответственно, $\operatorname{Re}g_\omega$ – регуляризация весов (определена ниже), $\operatorname{Re}g_\rho$ – регуляризация разреженности, определяемая как

$$\operatorname{Re}g_\rho=\sum_{g=1}^{G}\rho\log\frac{\rho}{\hat{\rho}_g}+(1-\rho)\log\frac{(1-\rho)}{\left(1-\hat{\rho}_g\right)},$$

где ρ – параметр разреженности,

$$\hat{\rho}_g=\frac{1}{M}\sum_{m=1}^{M}z_g(F_m)=\frac{1}{M}\sum_{m=1}^{M}h^{(1)}(\omega_g^{(1)}F_m+b_g^{(1)})\ -$$

среднее значение функции активации z_g нейрона g, G – количество скрытых нейронов.

$\operatorname{Re}g_\rho$ принимает нулевое значение, если $\rho=\hat{\rho}_g$, и увеличивается по мере их отклонения друг от друга.

Во время обучения сети значение $\operatorname{Re}g_\rho$ может стать маленьким ввиду уменьшения z. Данная проблема решается путем добавления члена регуляризации весов:

$$\operatorname{Re}g_\omega=\frac{1}{2}\sum_{x=1}^{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\omega_{mn}{}^{(x)}\right)^2,$$

где x – номер слоя.

1.2. Применение непрерывного вейвлет-преобразования и обнаружение аномалий

На основе непрерывного вейвлет-преобразования выполняется отображение функции $\hat{f}$ (см. (3)) в вейвлет-пространство:

$$W\hat{f}(s,u)=\int_{-\infty}^{+\infty}\hat{f}(t)\frac{1}{\sqrt{s}}\Psi^*\left(\frac{t-u}{s}\right)dt, \tag{4}$$

где Ψ – вейвлет, s – масштаб, u – сдвиг по времени, $s,u\in\Re$ ($\Re$ – действительные числа) $s\neq 0$.

Поскольку амплитуда вейвлет-коэффициентов $|W\hat{f}(s,u)|$ характеризует амплитуду локальной особенности функции на масштабе s в окрестности точки $t=u$ [13], возрастание амплитуды свидетельствует о возникновении аномалии в окрестности этой точки. В этом случае для обнаружения аномалий на масштабе s могут быть применены пороги T_s^l, которые с учетом изменения динамики процесса будем определять в скользящем временном окне:

$$P_{T_s^l}\left[W\hat{f}(s,u)\right]=\begin{cases}W\hat{f}(s,u), если\left|W\hat{f}(s,u)\right|\geq T_s^l,\\ 0, если\left|W\hat{f}(s,u)\right|<T_s^l,\end{cases} \tag{5}$$

где $T_s^l=q\times\sigma_s^l$, σ_s^l – среднеквадратическое отклонение (СКО) коэффициентов, рассчитанное в скользящем окне длины l, q – пороговый коэффициент.

Для многомасштабных аномалий их мощность в момент времени $t=u$ может быть определена как

$$E(u)=\sum_s P_{T_s^l}\left[W\hat{f}(s,u)\right], \tag{6}$$

которая будет положительной в случае повышений значений функции относительно характерного в рамках временного окна уровня (положительная аномалия) и будет отрицательной в случае понижений значений функции относительно характерного в рамках временного окна уровня (отрицательная аномалия).

Пороги разбивают пространство значений анализируемой функции на две непересекающиеся области Y_0 и Y_1. При использовании определенных порогов T_s^l (см. (5)) для заданного состояния данных η_r средняя величина потерь может быть оценена как

$$R_r(y)=\sum_{c=0}^{1}\Pi_{rc}P\{y\in Y_c/\eta_r\},$$

где Π_{rc} – функция потерь, $P\{y\in Y_c/\eta_r\}$ – условная вероятность попадания в область значений анализируемой функции Y_c, если в действительности имеет место состояние η_r, $r\neq c, r, c$ – индексы состояний (знак "/" означает условную вероятность).

Усредняя условную функцию риска по всем состояниям η_r для простой функции потерь

$$\Pi_{rc}=\begin{cases}1, r\neq c,\\ 0, r=c,\end{cases}$$

имеем апостериорный риск [23]:

$$R=\sum_{r\neq c}P\{\eta_r/y\in Y_c\},$$

где $P\{\eta_r/y\}$, $r=0,1$ – апостериорные вероятности.

Задача состоит в определении порогов, минимизирующих апостериорный риск:

$$R_{\min}=\min_{T_s^l\in Q_T}\sum_{r\neq c}P\{\eta_r/y\in Y_c\},$$

где Q_T – множество порогов.

Но учитывая неполные априорные знания о процессе, оцененные пороги могут всё же давать большое значение риска. В этом случае для минимизации потерь логично использовать разные факторы (наборы данных), характеризующие состояние исследуемого процесса, и принимать решение по совокупности. Имея полученные для каждого набора i значения величины E_i (см. (6)), логично применить простое правило: аномалия есть, если

$$\left|E_i\right| \geq T_{i,критич} \text{ для } i \in I, \tag{7}$$

где $T_{i,критичн}$ – наперед заданные критические значения, которые могут быть оценены для каждого набора i путем апостериорного риска [22], I – множество индексов набора данных, определяемое задачей исследования.

2. Применение метода для данных нейтронных мониторов

Используя эквивалентность непрерывного и дискретного вейвлет-преобразования, для выполнения операции (4) дискретный сигнал $\hat{f}[n]$ представлялся в виде ряда:

$$\hat{f}[n] = \sum_{j,k=-\infty}^{\infty} c_{jk} \Psi_{jk}[n], \tag{8}$$

где $\Psi_{jk} = 2^{j/2}\Psi(2^j n - k)$, $j,k \in \mathrm{N}$, $c_{jk} = \left\langle \hat{f}, \Psi^*_{jk} \right\rangle$ – коэффициенты разложения функции $\hat{f}$ в ряд по вейвлетам, которые определяются как:

$$c_{jk} = W\hat{f}\left(\frac{1}{2^j}, \frac{k}{2^j}\right).$$

Блок-схема реализации метода показана на рис. 1.

В экспериментах использовались данные сети станций нейтронных мониторов [8]. Данные каждой станции приняты за отдельный набор данных i, соответственно, нейронные сети строились отдельно для каждой станции. Также с учетом свойств динамики космических лучей анализ данных выполнялся отдельно для разных уровней солнечной активности. Следуя результатам работ [19 – 21], размерность входных векторов сети составляла 1440 отсчетов, что соответствует суткам (минутные данные). Операция (8) выполнялась с использованием вейвлетов Койфлет 2 [25]. При выполнении операции (5) использовались пороги $T_s^l = 2{,}5 \times \sigma_s^l$, длина скользящего временного окна $l = 1440$. Для обнаружения аномалий применялось правило (7), в котором значения порогов $T_{i,критичн} = 1{,}5 \times 10^3$ получены для каждой станции путём апостериорного риска. Мощность множества индексов $|I| = 2$.

На рис. 2 представлен пример применения метода к данным нейтронного монитора станции Оулу за 6 – 8 августа 2019 года. На рис. 2*а* серым цветом представлены исходные данные, черным – результат операции (3). Анализ результата на рис. 2*а* подтверждает эффективность применения нейронной сети для подавления шума. На рис. 2*б, г* представлены результаты операций (5), (6) к исходным данным нейтронного монитора, а на рис. 2*в, д* – к значениям функции $\hat{f}$ (после применения НС). Вертикальными линиями отмечены моменты регистрации аномалий 8 августа 2019 г.: начало слабой магнитной бури на высоких широтах в 07:00 [26] и Форбуш-эффекта в 14:00 мирового времени [27]. Результаты показывают, что применение операции (3) уменьшило влияние шумового фактора, связанного с суточным ходом вариаций космических лучей 6 августа (рис. 2*б, г*), и повысило эффективность обнаружения аномалии, возникшей в данных нейтронного монитора 7 – 8 августа за 14 часов до начала магнитной бури (рис. 2*в, д*).

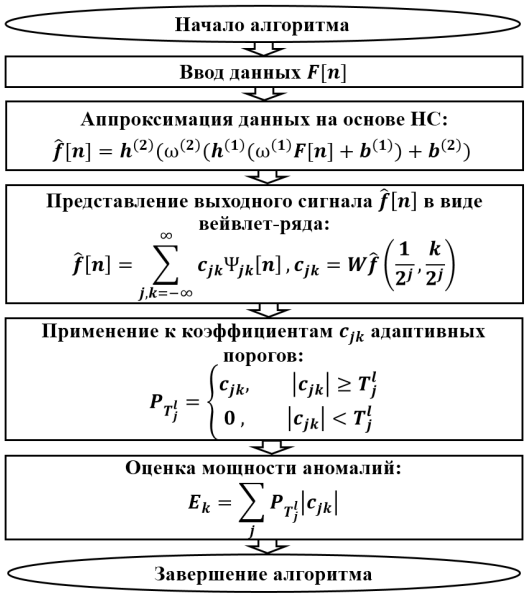

Рис. 1. Блок-схема алгоритма анализа данных сложной структуры

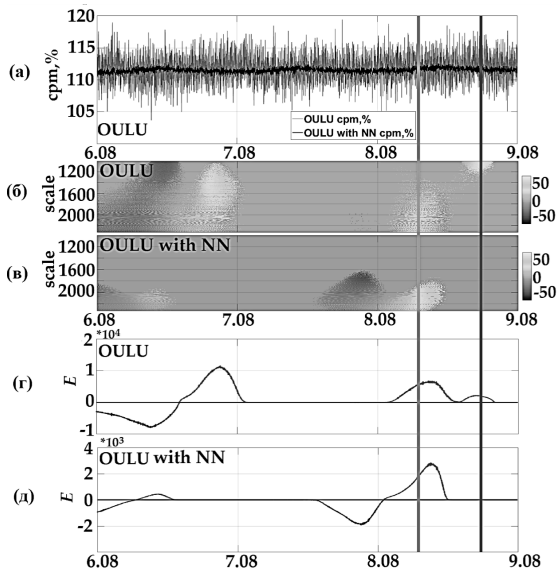

Рис. 2. Результаты обработки данных НМ за период с 6 по 8 августа 2019 г.

На рис. 3 представлен пример применения метода к данным сети станций нейтронных мониторов в период умеренной магнитной бури 10 – 11 сентября 2018 года. В верхней части рис. 3 для анализа состояния околоземного космического пространства (ОКП) показаны параметры скорости солнечного ветра (ССВ) (рис. 3*а*), данные межпланетного магнитного поля (Bz компонента) (рис. 3*б*) и данные Dst-индекса

геомагнитной активности (рис. 3*в*). Накануне события состояние ОКП было спокойное. По данным ИЗМИРАН [27], в конце суток 8 сентября в космических лучах был зарегистрирован Форбуш-эффект малой амплитуды, он отмечен на рис. 3 вертикальной линией. Результаты применения метода (рис. 3*г-и*) показывают в момент регистрации Форбуш-эффекта возникновение аномалий на всех анализируемых станциях – понижение интенсивности КЛ, имеющее наибольшую амплитуду на станции Оулу (рис. 3*д, з*). Далее, во второй половине суток 10 сентября, в связи с приходом неоднородного ускоренного потока от корональной дыры [26], ССВ повысилась (рис. 3*а*), флуктуации Bz компоненты ММП возросли (рис. 3*б*). Магнитная буря была зарегистрирована 10 сентября в 11:00 мирового времени ([26], отмечена вертикальной линией). Результаты обработки показывают до начала магнитной бури возникновение аномалии в данных нейтронных мониторов станций Инувик и Оулу (рис. 3*г, д, ж, з*). На станции Инувик порог E_i превысил критическое значение $T_{i,критичн}$ за 5 часов до начала магнитной бури, а на станции Оулу – за 1 час до начала магнитной бури. Значения $T_{i,критичн}$ отмечены на рис. 3*ж-и* горизонтальными линиями. Наиболее сильные изменения интенсивности космических лучей наблюдаются во время события (рис. 3*г, ж*), в период резкого возрастания ССВ (рис. 3*а*), увеличения флуктуаций ММП (рис. 3*б*) и существенного понижения Dst-индекса (рис. 3*в*). Результаты показывают сложную динамику космических лучей в возмущенный период и подтверждают высокую эффективность предлагаемого метода для обнаружения аномалий. Применение метода позволило детектировать аномальные изменения в данных нейтронных мониторов разных станций, возникшие накануне и в период события. По сравнению с методом ИЗМИРАН [27], предлагаемый метод позволил выполнить детальный анализ динамики космических лучей и обнаружить аномальное повышение их интенсивности за 1 – 5 часов до начала магнитной бури.

На рис. 4 представлен пример применения метода к данным нейтронных мониторов в период умеренной магнитной бури 5 августа 2019 г. В верхней части рис. 4 показаны параметры ССВ (рис. 4*а*), данные ММП (рис. 4*б*) и значения Dst-индекса геомагнитной активности (рис. 4*в*). По данным ИЗМИРАН [27], 4 августа зарегистрирован Форбуш-эффект малой амплитуды, он отмечен на рис. 4 вертикальной линией. В момент регистрации Форбуш-эффекта результаты обработки показывают возникновение аномалий на станциях Инувик (рис. 4*г, ж*) и Туле (рис. 4*е, и*). В начале суток 5 августа в период возрастания флуктуаций Bz компоненты ММП (рис. 4*б*) и резкого возрастания ССВ (рис. 4*а*) началась магнитная буря и одновременно зарегистрирован Форбуш-эффект в космических лучах [26, 27]. По результатам обработки (рис. 4*г-и*), в этот период на всех анализируемых станциях возникло аномальное повышение в космических лучах, имеющее наибольшую амплитуду на станции Инувик (рис. 4*г, ж*) и превысившее порог $T_{i,критичн}$ за 8 ч до начала магнитной бури. Результаты подтверждают сложную динамику космических лучей, изучение которой требует использования данных сети станций нейтронных мониторов и применения комплекса методов и подходов. Также результаты подтверждают эффективность предлагаемого метода для детального анализа данных космических лучей и обнаружения аномалий.

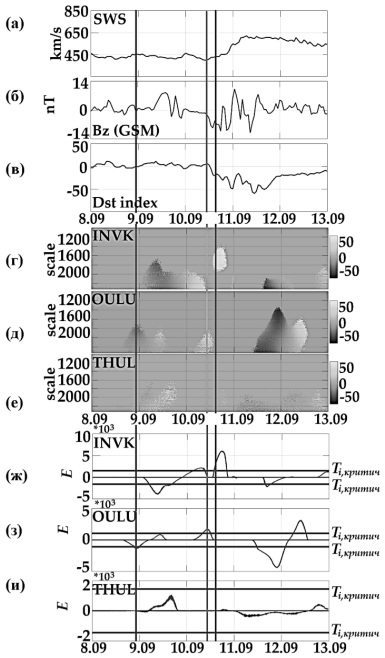

Рис. 3. Результаты обработки данных НМ за период с 8 по 14 сентября 2018 г.

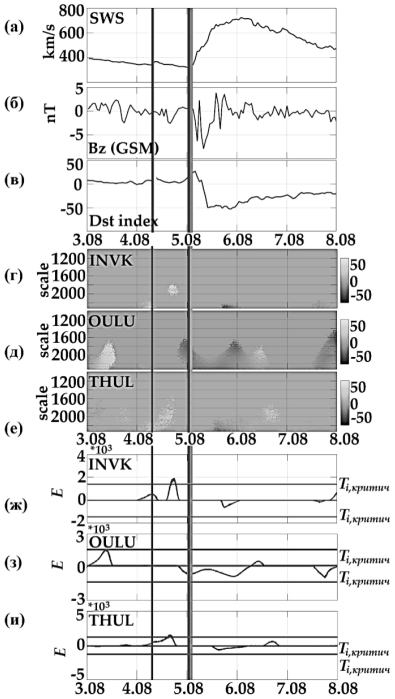

Рис. 4. Результаты обработки данных НМ за период с 3 по 8 августа 2019 г.

Для оценки эффективности метода выполнялась обработка данных высокоширотных станций нейтронных мониторов, которые вследствие особенностей анизотропии космических лучей менее подвержены влиянию шумовых факторов [9, 24]. Для верификации результатов использовался ресурс [26], содержащий базу данных Форбуш-эффектов и межпланетных возмущений. Результаты оценки представлены в табл. 1, которые показывают, что результативность метода превышает 82 %. Частота возникновения ложной тревоги составляет менее 10 % от общего числа событий.

Табл. 1. Результаты оценки эффективности метода

Год	Количество аномалий в сигнале	Результаты метода
2018	98	Выявлено: 86 %
		Не выявлено: 14 %
		Ложная тревога: 9 событий
2019	97	Выявлено: 84 %
		Не выявлено: 16 %
		Ложная тревога: 8 событий
2020	85	Выявлено: 82 %
		Не выявлено: 18 %
		Ложная тревога: 6 событий

Выводы

Предлагаемый метод позволяет изучить сложную структуру данных, подавить шум и детектировать аномалии разной формы и длительности. Применение эвристического подхода с элементами машинного обучения позволило оптимизировать параметры метода при наличии неполных знаний о структуре данных. Для адаптации предлагаемой нейросетевой модели предложено использовать регуляризаторы, минимизирующие ошибку нейронной сети на этапе её обучения. Путем оценки апостериорного риска предложен способ определения порогов, детектирующих аномалии в данных. Для минимизации потерь рекомендовано использовать наборы данных и принимать решение по совокупности.

Применение метода к данным сети станций нейтронных мониторов за период 2018–2020 гг. подтвердило его эффективность. Оценки показали, что результативность метода превышает 82 %, частота возникновения ложной тревоги составляет менее 10 %.

Сравнение предлагаемого метода с методом ИЗМИРАН [26] показало его эффективность в задаче обнаружения Форбуш-эффектов в космических лучах. Метод позволяет по данным сети станций нейтронных мониторов обнаружить Форбуш-эффекты, предшествующие началу магнитных бурь и служащие их предикторами. Численная реализация метода обеспечивает возможность его применения в оперативном анализе данных, что представляет интерес в задачах прогноза космической погоды и определяет прикладную значимость исследования.

Благодарности

Авторы выражают благодарность институтам, выполняющим поддержку станций нейтронных мониторов (www.nmdb.eu, http://spaceweather.izmiran.ru), которые использовались в работе.

Работа выполнена в рамках Государственного задания по теме «Физические процессы в системе ближнего космоса и геосфер при солнечных и литосферных воздействиях» (2021–2023 гг.), регистрационный номер АААА-А21-121011290003-0.

References

[1] Vorobiev AV, Vorobieva GR. Geographic information system for amplitude-frequency analysis of observation data of geomagnetic variations and space weather. Computer Optics 2017; 41(6): 963-972. DOI: 10.18287/2412-6179-2017-41-6-963-972.

[2] Mandrikova OV, Zhizhikina EA. An automatic method for assessing the state of the geomagnetic field. Computer Optics 2015; 39(3): 420-428. DOI: 10.18287/0134-2452-2015-39-3-420-428.

[3] Mandrikova OV, Stepanenko AA. An automated method for calculating the dst-index based on the wavelet model of the geomagnetic field variations field. Computer Optics 2020; 44(5): 797-808. DOI: 10.18287/2412-6179-CO-709.

[4] Mandrikova OV, Fetisova NV, Polozov YA. Hybrid Model for Time Series of Complex Structure with ARIMA Components. Mathematics 2021; 9: 1122. DOI: 10.3390/math9101122.

[5] Abunin AA, Abunina MA, Belov AV, Eroshenko EA, Oleneva VA, Jahnke V. Forbush effects with sudden and gradual onset [In Russian]. Geomagn Aeron 2012; 52(3): 313-320.

[6] Kuznetsov VD. Space weather and risks of space activity [In Russian]. Space Engineering and Technology 2014; 3(6): 3-13.

[7] Murzin BC. Astrophysics of cosmic rays: Textbook for universities [In Russian]. Moscow: “Logos” Publisher; 2007. ISBN: 978-5-98704-171-6.

[8] Real-time database of high-resolution neutron monitors. Source: ⟨www.nmdb.eu⟩.

[9] Belov AV, Eroshenko EA, Yanke VG, Oleneva VA, Abunina MA, Abunin AA. Method of global survey for the world network of neutron monitors [In Russian]. Geomagnetism and Aeronomy 2018; 58(3): 374-389. DOI: 10.7868/S0016794018030082.

[10] Mavromichalaki H, Souvatzoglou G, Sarlanis C, et al. Using real time neutron monitor database to establish an alert signal. Proc 31st Int Cosmic Ray Conference (ICRC) 2009. Source: ⟨https://galprop.stanford.edu/elibrary/icrc/2009/preliminary/pdf/icrc1381.pdf⟩.

[11] Veselovsky IS, Yakovchuk OS. On the forecast of solar proton events according to the data of ground-based neutron monitors [In Russian]. Astronomical Herald: Solar System Exploration 2011; 45(4): 365-375.

[12] Wawrzynczak A, Kopka P. Approximate Bayesian computation for estimating parameters of data-consistent forbush decrease model. Entropy 2018; 20: 622. DOI: 10.3390/e20080622.

[13] Chui CK. Introduction to wavelets [In Russian]. Moscow: "Mir" Publisher; 2001. ISBN: 5-03-003397-1.

[14] Astafieva NM. Wavelet analysis: basic theory and some applications. Physics–Uspekhi 1996; 39(11): 1085-1108. DOI: 10.1070/PU1996v039n11ABEH000177.

[15] Vizilter YuV, Gorbatsevich VS, Zheltov SYu. Structural and functional analysis and synthesis of deep convolutional neural networks. Computer Optics 2019; 43(5): 886-900. DOI: 10.18287/2412-6179-2019-43-5-886-900.
[16] Soldatova OP, Lezin IA, Lezina IV, Kupriyanov AV, Kirsh DV. Application of fuzzy neural networks to determine the type of crystal lattices observed in nanoscale images. Computer Optics 2015; 39(5): 787-794. DOI: 10.18287/0134-2452-2015-39-5-787-794.
[17] Rodin IA, Khonina SN, Serafimovich PG, Popov SB. Recognition of the types of wavefront aberrations corresponding to individual Zernike functions from the pattern of the point scattering function in the focal plane using neural networks. Computer Optics 2020; 44(6): 923-930. DOI: 10.18287/2412-6179-CO-810.
[18] Goodfellow Y, Benjio I, Courville A. Deep learning. Adaptive computation and machine learning series. Cambridge, London: The MIT Press; 2016.
[19] Geppener VV, Mandrikova BS. An automated method for analyzing cosmic ray data and isolating sporadic effects [In Russian]. Computational Mathematics and Mathematical Physics 2021; 61(7): 1137-1148. DOI: 10.31857/S0044466921070061.
[20] Mandrikova O, Mandrikova B, Rodomanskay A. Method of constructing a nonlinear approximating scheme of a complex signal: Application pattern recognition. Mathematics 2021; 9(7): 737. DOI: 10.3390/math9070737.
[21] Geppener VV, Mandrikova B. Detecting and identifying anomalous effects in complex signals. Autom Remote Control 2021; 82(10), 1668-1678. DOI: 10.1134/S0005117921100052.
[22] Wald A. Statistical decision functions. London: Chapman & Hall; 1950.
[23] Bansal AK. Bayesian parametric inference. Oxford, UK: Alpha Science International Ltd; 2007.
[24] Abunina MA. Anisotropy of cosmic rays in various structures of the solar wind [In Russian]. The thesis for the Candidate's degree in Technical Sciences. Moscow; 2016.
[25] Daubechies I. Ten lectures on wavelets [In Russian]. Moscow: "RKhD"Publisher; 2001.
[26] IZMIRAN catalog of Forbush effects and interplanetary disturbances [In Russian]. Source: ⟨http://spaceweather.izmiran.ru/rus/fds2019.html⟩.
[27] Institute of Applied Geophysics [In Russian]. Source: ⟨http://ipg.geospace.ru/⟩.

Сведения об авторе

Мандрикова Богдана Сергеевна, 1996 года рождения. В 2018 году окончила Белгородский государственный технологический университет имени В.Г. Шухова по специальности 09.03.01 «Информатика и вычислительная техника». В 2020 году с отличием окончила Санкт-Петербургский электротехнический университет «ЛЭТИ» по специальности 01.04.02 «Прикладная математика и информатика». Работает младшим научным сотрудником лаборатории системного анализа Института космофизических исследований и распространения радиоволн ДВО РАН. Область научных интересов: вейвлеты, нейронные сети, анализ сложных сигналов, исследование данных нейтронных мониторов и обнаружение аномалий. Автор более 20 публикаций (статей).
E-mail: *555bs5@mail.ru* .

ГРНТИ: 27.41.17
Поступила в редакцию 19 декабря 2021 г. Окончательный вариант – 6 февраля 2022 г.

Дизайн: М.А. Вахе, А.А. Алексеев. Оформление и вёрстка: М.А. Вахе, С.В. Смагин, И.А. Кондратьев.
Лит. редактор и корректор Ю.Н. Литвинова.
Консультант по оформлению англоязычного блока М.И. Котляр, консультант по оформлению литературы Е.В. Семиколенных.
E-mail: *journal@computeroptics.ru*, *http://www.computeroptics.ru*

Подписано в печать 04.05.2022 г. Усл. печ. л. 19,1.
Заказ № 11/6. Тираж 206 экз. Печать офсетная. Формат 62×84 1/8.
Цена: 550 рублей / Price of 550 rubles (6+)

Издатель: Институт систем обработки изображений РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН, (443010, г. Самара, ул. Молодогвардейская, 151)
Учредители: Федеральное государственное автономное образовательное учреждение высшего образования
«Самарский национальный исследовательский университет имени академика С.П. Королева» (443086, г. Самара, Московское шоссе, д.34),
Федеральное государственное учреждение «Федеральный научно-исследовательский центр «Кристаллография и фотоника» Российской академии наук» (117342, г. Москва, ул. Бутлерова, д 17А)
Отпечатано в типографии ООО «Предприятие «Новая техника» (443013 г. Самара, пр-кт. Карла Маркса, 24-76)

A method for analyzing complex structured data with elements of machine learning

B.S. Mandrikova [1]

[1] *Institute of Cosmophysical Research and Radio Wave Propagation, Far Eastern Branch of the Russian Academy of Sciences, 684034, Kamchatskiy Kray, Paratunka, Russia, Mirnaya st, 7*

Abstract

A method for analyzing data of complex structure based on combining a wavelet transform and neural networks Autoencoder is proposed. The method allows you to research the data structure, detect abnormal changes of various shapes and durations, and suppress noise. The efficiency of the method is shown on the example of data from a network of neutron monitor stations. Neutron monitor data determine the intensity of secondary cosmic rays and are one of the key factors in space weather. The numerical implementation of the method allows it to be applied on-line, which is of interest in problems of analyzing environmental data and detecting catastrophic events.

Keywords: data analysis, data of complex structure, wavelet transform, neural networks, neutron monitors.

Citation: Mandrikova BS. A method for analyzing complex structured data with elements of machine learning. Computer Optics 2022; 46(3): 506-512. DOI: 10.18287/2412-6179-CO-1088.

Acknowledgements: The work was funded under the government project AAAA-A21-121011290003-0 “Physical processes in the system of near space and geospheres under solar and lithospheric influences” IKIR FEB RAS.

Author’s information

Bogdana Sergeevna Mandrikova, (b. 1996), graduated from the Belgorod State Technological University named after V.G. Shukhov in 2018. Graduated with honors from the Saint Petersburg Electrotechnical University 'LETI' in 2020. Junior researcher of the Laboratory of System Analysis at the Institute of Cosmophysical Research and Radio Wave Propagation Far East Branch Russian Academy of Sciences. Her scientific interests are wavelets, neural networks, analysis of complex signals, research of neutron monitors data and detect anomalies. Author of over 20 publications (articles). E-mail: *555bs5@mail.ru* .

Received December 19, 2021. The final version – February 06, 2022.