УДК 004.85 : 551.50 DOI: 10.18698/2309-3684-2021-3-88104

Методы интеллектуального анализа данных в модели наукастинга опасных явлений

© А.О. Шершакова[1], В.П. Пархоменко[1,2]

[1]МГТУ им. Н.Э. Баумана, Москва, 105005, Россия
[2]Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН
Москва, 119333, Россия

Настоящая работа посвящена исследованию и применению методов интеллектуального анализа для реализации схемы наукастинга опасных явлений. В ходе работы были сформированы большие наборы данных на основе метеорологических наблюдений облачных ячеек, отличающиеся методами обработки информации для их подготовки. Для каждого набора был построен ряд математических моделей классификации облачных ячеек по степени опасности формирования из них смерчей. В качестве основного языка разработки выбран язык программирования Python. Работа имеет большое практическое значение в сфере прогнозирования погодных явлений. Ее новизна заключается в использовании современной методологии машинного обучения вместо традиционного подхода экстраполяции данных, широко применяемого в различных схемах наукастинга.

Ключевые слова: *наука о данных, интеллектуальный анализ данных, машинное обучение, математическая модель, Python, наукастинг, смерчи, облачные ячейки*

Введение. Термин наукастинг определяется как прогнозирование погодных явлений с местной детализацией любым методом, на период от настоящего времени до 6 часов вперед, включая подробное описание текущей погоды [1]. Наукастинг обычно применяется к прогнозу мезомасштабных или локальных явлений и дается на довольно короткий промежуток времени (несколько часов). Основной задачей наукастинга является выявление признаков образования конвективных облаков, генерирующих опасные явления, отслеживание их перемещения и прогнозирование возможного развития [2, 3].

Целью данной работы является исследование методов интеллектуального анализа данных и машинного обучения, разработка эффективных алгоритмов их применения для реализации схем наукастинга. Интеграция методов машинного обучения с классическими методами анализа может способствовать выявлению скрытых особенностей и закономерностей протекания атмосферного процесса, учет которых в перспективе позволит повысить качество его прогнозирования.

Анализ и этапы подготовки данных. Анализ данных — совокупность всех методов работы с информацией, позволяющих получать новые данные. Анализ покрывает стадии сбора, подготовки, исследования данных, толкования результатов. Теоретические методы

анализа данных лежат в основе алгоритмов машинного обучения, применяемого на стадии моделирования. Большинство методов анализа имеют статистическую теоретическую базу, для выявления свойств изучаемых объектов ставятся такие классические задачи математической статистики, как нахождение характеристик случайных величин, выявление корреляционных зависимостей в данных [4, 5].

Математическое и численное моделирование является важным инструментом для исследования сложных систем. Современное моделирование проводится с помощью развитых программных средств, в том числе отечественных, например, для решения задач нестационарной газодинамики многокомпонентного газа различными численными методами [6–9], крупномасштабных процессов в окружающей среде [10].

На практике часто можно столкнуться с большой размерностью данных, которая вносит некоторые трудности в обработку информации. Один из известных методов анализа данных, использующийся для снижения размерности и избавления от корреляционных зависимостей — метод главных компонент. Идея метода состоит в преобразовании множества объясняющих переменных в новое меньшее множество попарно некоррелированных переменных, по которым можно достаточно точно воспроизвести важные свойства анализируемого массива данных [11].

Пусть имеется множество исходных числовых признаков $f_1(x),\ldots, f_n(x)$, необходимо перейти в пространство новых числовых признаков $g_1(x),\ldots,g_m(x),\ m \le n$. Формулируется требование о возможности линейного восстановления старых признаков по новым:

$$\hat{f}_j(x) = \sum_{k=1}^{m} g_k(x) u_{jk},\ \ j = 1,\ldots,n.,\ \ \forall x \in X,$$

где $\hat{f}_j(x)$ — восстановленные признаки, $g_k(x)$ — новые признаки, u_{jk} — матрица преобразований, восстанавливающая старые признаки по новым.

При этом:

$$\sum_{i=1}^{l}\sum_{j=1}^{n}(\hat{f}_j(x_i) - f_j(x_i))^2 \to \min_{\{g_k(x_i)\},\{u_{jk}\}},$$

где $x_1,\ldots,x_l$ — выборка.

В матричном виде постановка задачи формулируется таким образом: имеются матрицы «объекты – признаки» . Строки — это объекты выборки, столбцы — это признаки (для F — старые, для G — новые):

$$F_{l\times n}=\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_l) & \dots & f_n(x_l) \end{pmatrix};\ G_{l\times m}=\begin{pmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \dots & \dots & \dots \\ g_1(x_l) & \dots & g_m(x_l) \end{pmatrix},$$

где $F_{l\times n}$ — старая матрица; $f_1(x),\dots, f_n(x)$ — старые признаки; $x_1,\dots,x_l$ — выборка; $G_{l\times m}$ — новая матрица; $g_1(x),\dots,g_m(x)$ — новые признаки.

Введем матрицу линейного преобразования перехода от новых признаков к восстановленным старым:

$$\underset{n\times n}{U}=\begin{pmatrix} u_{11} & \dots & u_{1m} \\ \dots & \dots & \dots \\ u_{n1} & \dots & u_{nm} \end{pmatrix}.$$

Требуется построить матрицы G и U таким образом, что бы их произведение в некотором смысле хорошо восстанавливало матрицу F:

$$\hat{F}=GU^T\approx F,$$

где $\hat{F}$ — восстановленная матрица F.

В матричном виде эта задача заключается в том, чтобы минимизировать квадрат нормы разности двух матриц:

$$\sum_{i=1}^{l}\sum_{j=1}^{n}(\hat{f}_j(x_i)-f_j(x_i))^2=\left\|GU^T-F\right\|^2\to\min_{G,U},$$

Если $m\le rang(F)$, то минимум $\left\|GU^T-F\right\|^2$ достигается, когда столбцы матрицы U определяются собственными векторами матрицы F^TF, соответствующими m максимальным собственным значениям $\lambda_1,\dots,\lambda_m$, а матрица $G=FU$.

Отдельной задачей является выбор числа главных компонент. Эффективная размерность m выборки E_m, определяется следующим образом: собственные значения матрицы FF^T упорядочиваются по убыванию, далее строится ряд, в котором выявляется наличие крутого склона или резкого обрыва, отличающие большие собственные значения от маленьких. Присутствие такого эффекта демонстрирует существоваие закономерности, благодаря которой данные образовали в $n-$мерном пространстве некое линейное подпространство меньшей размерности. Такой способ называют критерием «крутого склона»: $m: E_{m-1}\ge E_m$:

$$E_m = \frac{\left\| GU^T - F \right\|^2}{\left\| F \right\|^2} = \frac{\lambda_{m+1} + \ldots + \lambda_n}{\lambda_1 + \ldots + \lambda_n} \le \varepsilon. \qquad (1)$$

Оценить m можно также при помощи величины обратной к (1), которая называется долей объясненной дисперсии и равна отношению выборочной дисперсии $\sum_{j=1}^{n} \lambda_j$ к остаточной дисперсии $\sum_{j=m+1}^{n} \lambda_j$.

Перед анализом для корректной работы всех моделей данные обычно необходимо центрировать и нормировать в соответствии с формулой:

$$x_i^{j'} = \frac{x_i^j - \frac{1}{l}\sum_{i=1}^{l} x_i^j}{\sqrt{\frac{1}{l}\sum_{i=1}^{l} (x_i^j - \frac{1}{l}\sum_{i=1}^{l} x_i^j)^2}},$$

где $x_i^{j'}$ — преобразованные признаки; $\frac{1}{l}\sum_{i=1}^{l} x_i^j$ — выборочное среднее; $\sqrt{\frac{1}{l}\sum_{i=1}^{l} (x_i^j - \frac{1}{l}\sum_{i=1}^{l} x_i^j)^2}$ — выборочное среднеквадратичное отклонение. Объединение двух этих методов называют стандартизацией, она помогает избавиться от сдвигов и разниц в масштабах у признаков.

В вычислительной части работы была подтверждена некорректная работа метода главных компонент при отсутствии масштабирования (стандартизации). Большое различие в диапазонах признаков также может негативно сказываться при построении линейных моделей.

Постановка задачи машинного обучения для размеченных данных. В данной работе рассматривается машинное обучение с учителем, этот тип обучения предполагает наличие полного набора размеченных данных для тренировки модели на всех этапах ее построения [12–14]. Каждому примеру в обучающем наборе соответствует ответ, который должен получиться в результате работы алгоритма. В задачах обучения с учителем происходит восстановление общей закономерности по конечному числу примеров. Таким образом, решается задача методом интерполяции.

Далее используются следующие обозначения: x — объект; X — пространство объектов; $y = y(x)$ — ответ на объект x; Y — пространство ответов.

Объектом называется то, для чего нужно сделать предсказание. Пространство объектов — это множество всех возможных объектов, для которых может потребоваться делать предсказание. Ответом будем называть то, что нужно предсказать.

Пусть у нас имеется некоторая функция (алгоритм), отображающая множество объектов в множество ответов:

$$a(x): X \to Y. \tag{2}$$

Эта функция неизвестна, но определена на конечном множестве точек, которое называется «обучающей выборкой»:

$$X^l = (x_i, y_i)_{i=1}^l. \tag{3}$$

Таким образом, имеется l пар объект–ответ, по которым необходимо восстановить зависимость или построить каким-либо методом обучения функцию (2), которая бы аппроксимировала неизвестную зависимость. Обучающая выборка (3) — это примеры, на основе которых строится общая закономерность. Наиболее распространённый способ задания объектов — это признаковое описание:

$$f = (f^1(x), f^2(x), \ldots, f^n(x)).$$

Признаки — это функции, ставящие в соответствие объекту некоторые значения, как правило, числовые. В машинном обучении введено понятие матрицы «объекты–признаки»:

$$F = \left\| f_j(x_i) \right\|_{l \times n} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_l) & \ldots & f_n(x_l) \end{pmatrix}.$$

Важный принцип машинного обучения, общий для всех задач — это разделение решения на два этапа: обучение и тестирование (train и test). При этом выборка разбивается на две части. Вначале на этапе обучения $\mu: (X \times Y)^l \to A$ по первой части выборки $X^l = (x_i, y_i)_{i=1}^l$ строится алгоритм $a = \mu(X^l)$:

$$\begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_l) & \ldots & f_n(x_l) \end{pmatrix} \xrightarrow{y} \begin{pmatrix} y_1 \\ \ldots \\ y_l \end{pmatrix} \xrightarrow{\mu} a.$$

На этапе тестирования алгоритм $a = \mu(X^l)$ для оставшейся части обучающей выборки (тестовой выборке) из новых объектов $x_1', \ldots, x_k'$ выдает ответы (прогноз) $a(x_i')$:

$$\begin{pmatrix} f_1(x_1') & \ldots & f_n(x_1') \\ \ldots & \ldots & \ldots \\ f_1(x_k') & \ldots & f_n(x_k') \end{pmatrix} \xrightarrow{a} \begin{pmatrix} a(x_1') \\ \ldots \\ a(x_k') \end{pmatrix}.$$

Постановка задачи наукастинга. Сбор и подготовка данных. Основной задачей работы является изучение особенностей и закономерностей развития водяного смерча на некоторой территории с применением интеллектуального анализа данных и машинного обучения для наукастинга этого природного явления [15–18]. Исследуются возможные применения ряда математических моделей классификации облачных ячеек по степени опасности формирования из них смерчей на базе различных методов интеллектуального анализа данных:

- линейная классификация;
- логистическая регрессия;
- метод решающих деревьев;
- метрическая классификация.

Исходными данными для построения моделей являются:

- набор пространственных и временных данных с характеристиками облачных ячеек;
- перечень ячеек, из которых возникали смерчи.

Характеристики ячеек получены по результатам обработки спутниковых данных (со спутника Meteosat Second Generation — 10 (MSG3)). Спутник сканирует поверхность Земли каждые 15 минут. Собранная информация покрывает измерение параметров исследуемых объектов за несколько лет. Характеристики ячеек содержатся в текстовых файлах. Ячейки идентифицируются по двум полям — порядковый номер ячейки в пределах текущего цикла сканирования и дата сканирования. Ячейки прослеживаются во времени — устанавливается связь между ячейками на различных циклах сканирования, выявляется номер и дата сканирования так называемого родителя (на предыдущем цикле), если он существует [15, 19].

В качестве основного языка разработки выбран язык программирования Python, используются библиотеки Pandas, Matplotlib, Seaborn, NumPy**,** SciPy, Scikit-learn для решения задач интеллектуального анализа данных.

На первом этапе работы с данными из многочисленных файлов табличного формата, в которых находились значения параметров облачных ячеек, была получена информативная таблица, имеющая 349147 строк и 79 столбцов, строки соответствуют облачным ячейкам, а столбцы — их характеристикам. Отдельно подготовлена еще одна таблица, в ней перечислены все облачные ячейки (87 штук), которые реализовались в смерч.

Известно, что объекты прослеживаются во времени: ячейки идентифицируются номером ячейки в текущем цикле сканирования и датой сканирования, также для каждой ячейки указывается номер и дата сканирования ее предка (родителя) на предыдущем цикле, если

таковой имеется. Путем соотнесения этой информации были составлены истории для всех смерчевых ячеек, сформирована отдельная структура, необходимая на этапе исследования данных. На данных историй была построена был построен график наличия числа измерений для ячеек (рис. 1).

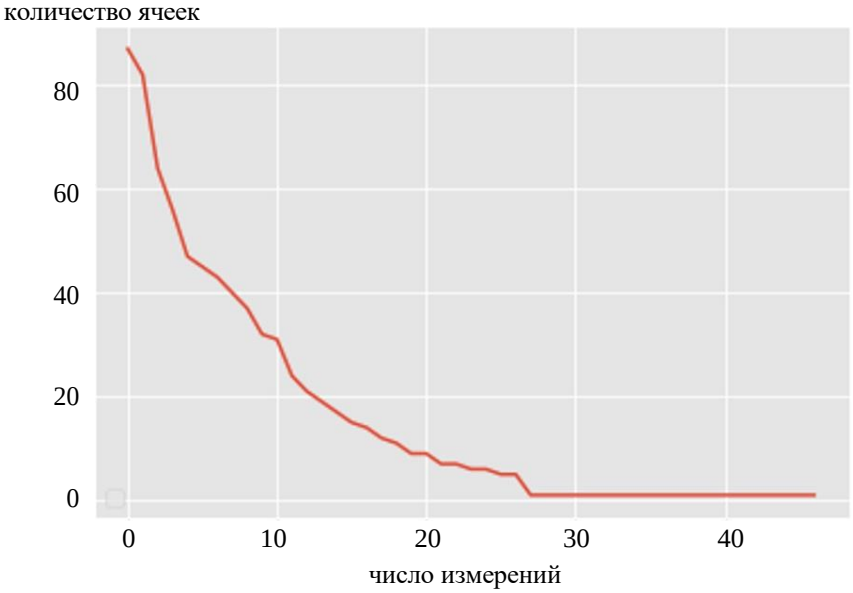


Рис. 1. График зависимости количества измерений для разных ячеек

После подготовки основных структур был реализован этап очистки данных: удалены неинформативные по отношению к задаче прогнозирования опасного явления характеристики, найдены и обработаны пропущенные значения.

Важным шагом в очистке данных является анализ информативности параметров. Программным путем были определены все характеристики, у которых более 95% значений одинаковы. Характеристики с полным отсутствием разнообразия были удалены, а остальные спорные по значимости параметры были проанализированы статистическими методами. Были получены основные сведения о значениях характеристик ячеек, построены гистограммы распределений значений всех параметров и диаграммы их размаха. Также было решено понаблюдать за изменением этих характеристик опасных (смерчевых и смерчеопасных) ячеек с течением времени, для этого были построены графики зависимости параметров от времени приближения к моменту реализации смерча.

В результате описанных действий были удалены неинформативные параметры и найдены параметры с особыми значениями на наборе со смерчевыми ячейками.

Во время выявления неинформативных данных были обнаружены повторяющиеся значения (выбросы), которые существенно

отличались от большей части данных и выходили за рамки физически допустимых. Был определен особый тип выбросов — код пустого значения, задаваемый, когда параметр по какой-то причине (например, недостаточно данных) не мог быть рассчитан. При помощи экспертных данных были определены допустимые значения параметров, все выбросы были удалены.

Анализ корреляционных зависимостей. Формирование обучающей выборки. После очистки данных от неинформативных значений и выбросов приступают к дополнительной стадии очистки: этап анализа корреляционных зависимостей признаков. Проводится стандартизацию признаков, так как коэффициент корреляции эффективно вычисляется только на данных с распределением близким к нормальному. Стандартизация подходит и для снижения разницы в масштабах, так как является методом шкалирования.

Рассматриваются два способа очистки от корреляционных зависимостей: удаление взаимозаменяемых параметров (УК–удаление коррелируемых параметров) и метод главных компонент (МГК — метод главных компонент). Независимое применение методов на одинаковом наборе данных позволит в дальнейшем провести сравнение и выявить оптимальный подход обработки информации.

Первый метод представляет собой отсеивание наиболее легко заменимых параметров, у которых имеется более одной пары с корреляцией 90% и выше. Была составлена таблица попарных корреляционных зависимостей, она была отсортирована по возрастанию абсолютной величины корреляции параметра со столбцом ответов, классифицирующим ячейки на опасные и безопасные по историческим данным. После чего происходило непосредственное удаление параметров с большой взаимной корреляционной зависимостью, при этом определенные сортировкой важные параметры были оставлены. Таким образом, из 79 признаков осталось 38 значимых, независимых от других характеристик.

Второй метод работы с корреляционными зависимостями заключался в привлечении метода главных компонент. Для шкалированных данных был построен график зависимости доли совокупной объясненной дисперсии от числа главных компонент (рис. 2), из него следует, что первые 15 компонент содержат примерно 85% дисперсии, а 25 компонент будет достаточно, чтобы покрыть 95% всего разнообразия. С помощью машинного обучения был построен график зависимости точности предсказания модели методом логистической регрессии от числа компонент (рис. 3). В результате было решено сформировать таблицу из 17 главных компонент, так как по данным рис. 2 и 3 именно 17 компонент дают максимальную среднюю точность логистической регрессии, при этом покрывается 90% совокупной объясненной дисперсии.

Таким образом, было получено два набора различной размерности, в которых минимизированы корреляционные зависимости параметров. Следующий шаг — формирование обучающей выборки.

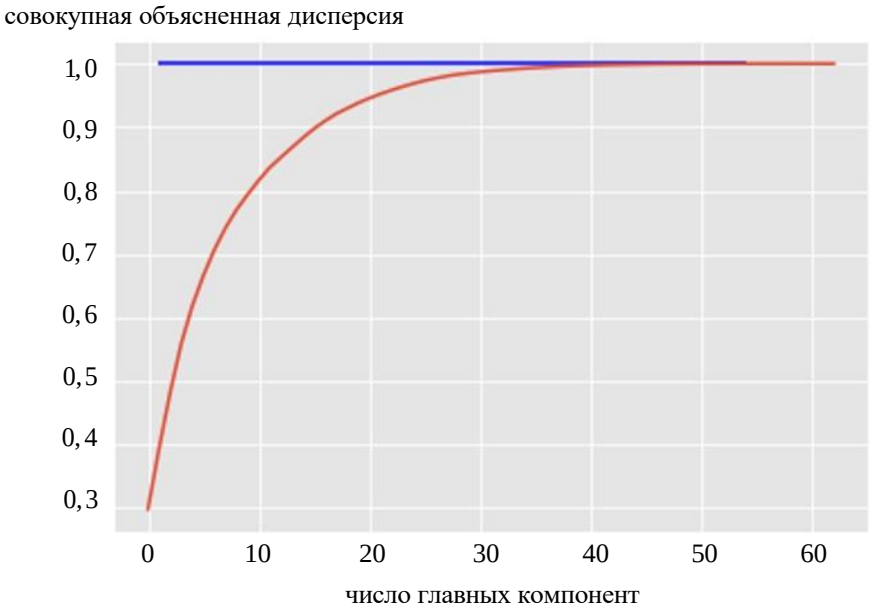


Рис. 2. График зависимости доли совокупной объясненной дисперсии от числа главных компонент

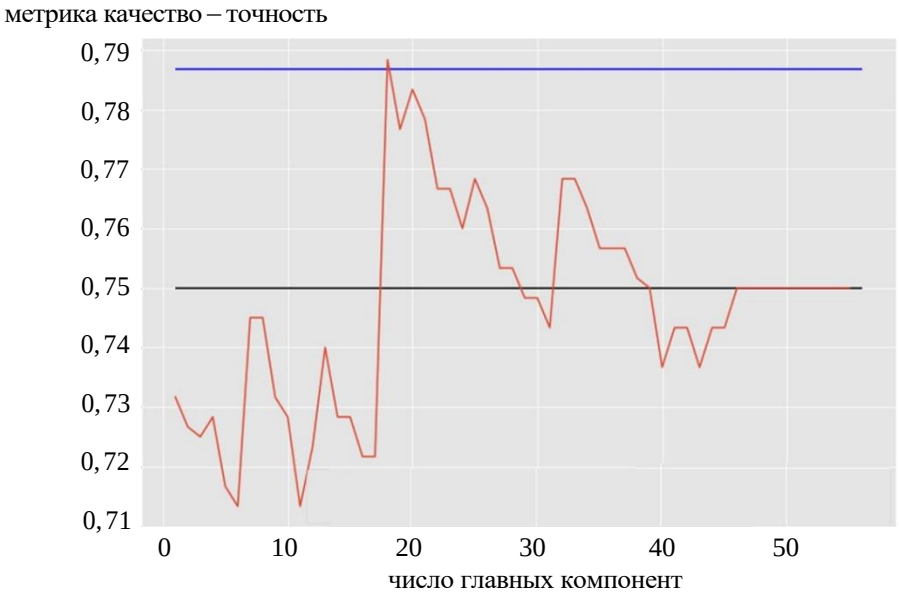


Рис. 3. График зависимости точности предсказания:
— — зависимость точности от числа компонент; — — средняя точность работы с корреляционными зависимостями; — — средняя точность собственного метода избавления от корреляций на данных без выбросов

Для того чтобы модель могла отличать смерчи от условно безопасных ячеек, необходимо чтобы обучающая выборка содержала соизмеримое число элементов классов (опасных и безопасных ячеек). На практике до очистки имелось 349060 представителей класса безопасных ячеек и всего 87 класса опасных.

Восстановление баланса классов может проходить двумя путями: удалением некоторого количества примеров преобладающего класса (undersampling) или увеличением количества примеров недостающей части (oversampling). В данной работе использовался undersampling, так как применение oversampling в чистом виде оказалось очень проблематичным, сгенерированные данные о смерчевых ячейках нарушали целостность исследования, вносили неопределённость в физическое соответствие данных и опасного явления (эти трудности были выявлены опытным путем).

При помощи моделей логистической регрессии и кросс–валидации было выявлено оптимальное соотношение опасных и безопасных ячеек для undersampling. По результатам анализа было подготовлено несколько выборок с соотношениями: 0,70; 0,75; 0,80; 0,85; 0,90; 0,95; 1,00.

Помимо методов восстановления баланса, которые не учитывают физические особенности явления, было приято решение ввести дополнительную характеристику, которая будет использоваться для выбора ячеек в обучающий набор. Новый параметр представляет собой расстояние от смерча в день его реализации и имеет вероятностный характер. Чем дальше рассматриваемая ячейка от смерчеобразующей, тем меньше вероятность p ее реализации:

- $p = 0,9$, если расстояние меньше или равно 50 км;
- $p = 0,8$, если расстояние от 50 до 100 км;
- $p = 0,7$, если расстояние от 100 до 200 км;
- $p = 0,6$, если расстояние от 200 до 300 км;
- $p = 0,5$, если расстояние от 300 до 400 км;
- $p = 0,4$, если расстояние от 400 до 500 км;
- $p = 0,3$, если расстояние от 500 до 600 км;
- $p = 0,2$, если расстояние от 600 до 700 км;
- $p = 0,1$, если расстояние от 700 до 800 км;
- $p = 0$, если расстояние больше 800 км.

На рис. 4 изображена раскрашенная карта ячеек в соответствии с рассчитанной вероятностной характеристикой для одной из опасных ячеек. Вероятностная интерпретация имеет ряд преимуществ и в дальнейшем может использоваться для построения более сложных прогностических моделей и реализации многоклассового разбиения

данных. Были составлены также таблицы: с опасными ячейками и максимально удаленными безопасными, с опасными ячейками и ближайшими безопасными ячейками для таблицы с выбросами. Из них были получены соизмеримые выборки по методу undersampling.

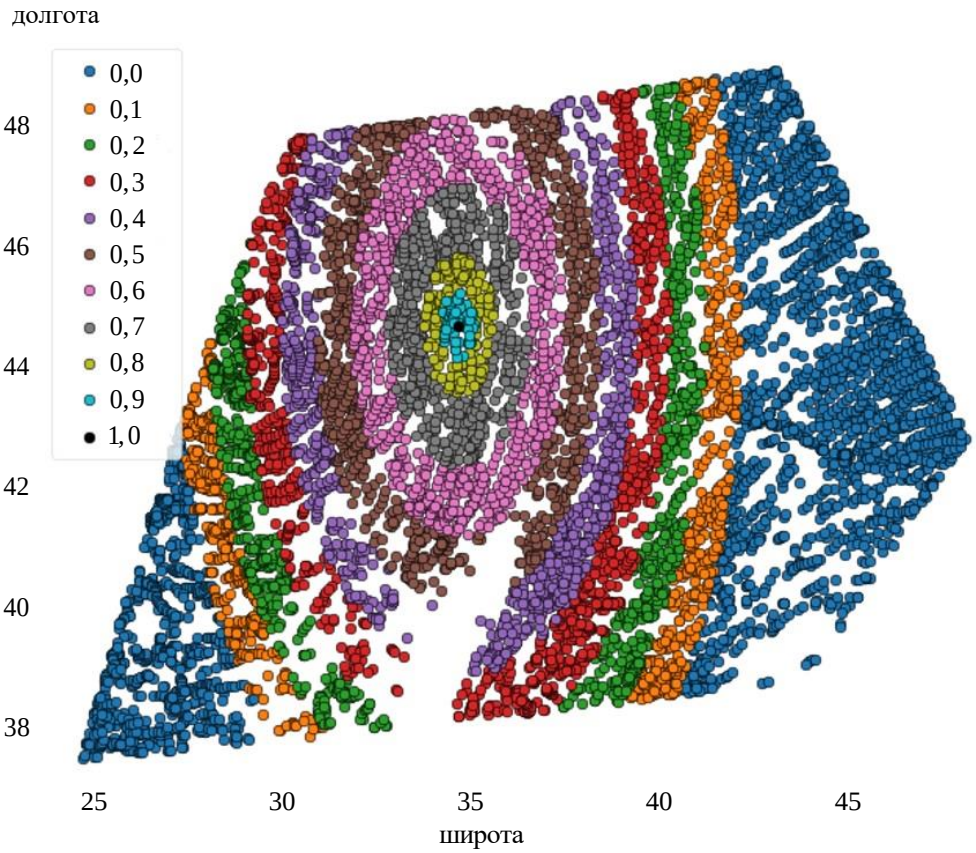


Рис. 4. Визуализация вероятностной характеристики расстояния от смерча в день его реализации для одной смерчевой ячейки

Моделирование и визуализация результатов. Моделирование производилось при помощи библиотеки Scikit-learn с использованием моделей линейной классификации, логистической регрессии, решающих деревьев, ближайшего соседа. Метрикой качества моделей была выбрана доля правильных ответов. С помощью кросс-валидации определена средняя точность для разных моделей, методов обаботки и разбиения данных. Из рис. 5 следует, что лучшие результаты при сравнении обработки корреляционных зависимостей на всех моделях показал метод отсеивания (УК).

Для сравнения работы моделей был построен совмещенный график для данных с отсеянными корреляциями (УК) (рис.6). Видно, что при соотношении классов 0,95 наиболее качественное разбиение было осуществлено при помощи модели линеной классификации.

Основываясь на этих результатах, можно сделать вывод о том, что признаки обладают свойством линейной сепарабельности.

Графики, представленные на рис. 7 демонстрируют чувствительность моделей как на далеких от смерча ячейках, так и на ближайших к нему. Лучшим образом на выборках, сформированных с помощью вероятностной характеристики, показала себя метрическая модель.

Рис. 5. Графики зависимости точности моделей от соотношений классов: *а* — — — метрический классификатор, с удаленными параметрами при анализе корреляций (УК), — — метрический классификатор, метод главных компонент (МГК); *б* — — — линейный классификатор (УК), — — линейных классификатор (МГК); *в* — — — логистическая регрессия (УК), — — логистическая регрессия (МГК); *г* — — — решающее дерево (УК), — — решающее дерево (МГК)

Для сравнения работы моделей был построен совмещенный график для данных с отсеянными корреляциями (УК) (рис.6). Видно, что при соотношении классов 0,95 наиболее качественное разбиение было осуществлено при помощи модели линеной классификации. Основываясь на этих результатах, можно сделать вывод о том, что признаки обладают свойством линейной сепарабельности.

Графики, представленные на рис. 7 демонстрируют чувствительность моделей как на далеких от смерча ячейках, так и на ближайших к нему. Лучшим образом на выборках, сформированных с помощью вероятностной характеристики, показала себя метрическая модель.

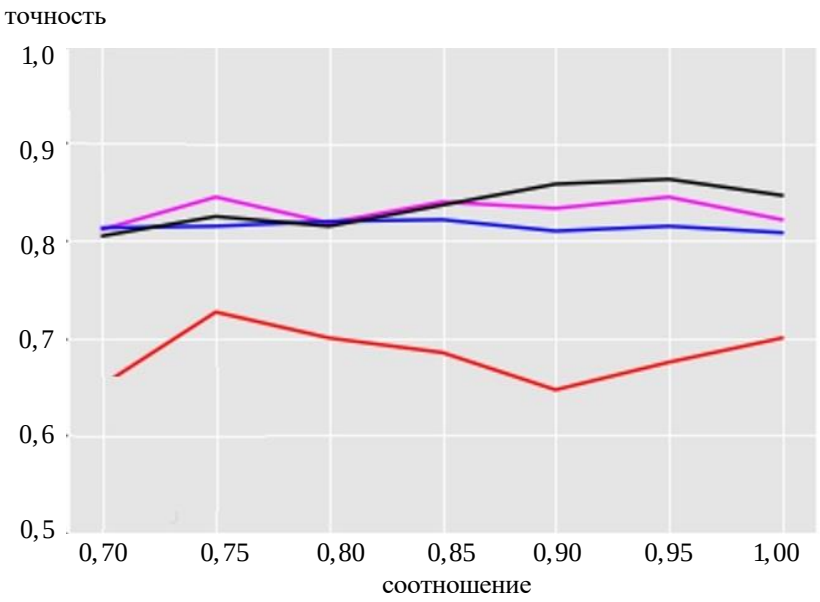
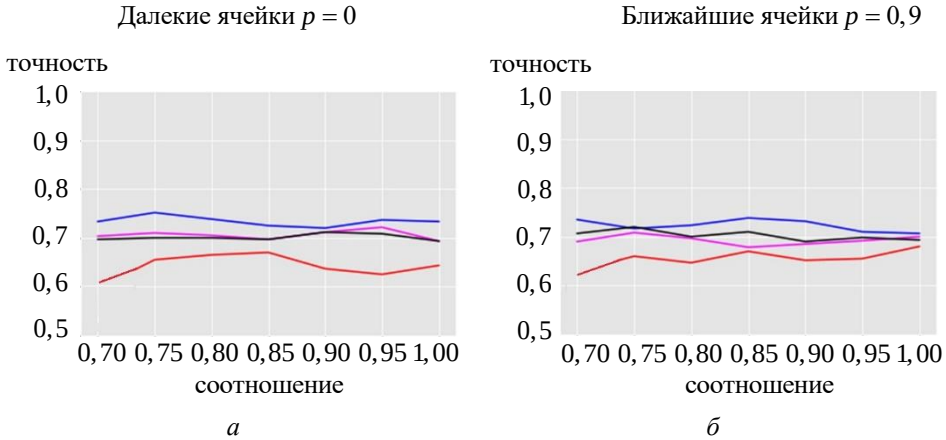


Рис. 6. Сравнение точности моделей (УК):
— — решающее дерево, — — логистическая регрессия,
— — метрический классификатор, — — линейная классификация



а *б*

Рис. 7. Точность классификатора для обучающих выборок, построенных для минимума и максимума вероятностной характеристики:
а — далекие ячейки $p = 0$; *а* — ближайшие ячейки $p = 0,9$:
— — решающее дерево, — — логистическая регрессия,
— — метрический классификатор, — — линейная классификация

Выводы. В работе была проведена качественная подготовка д анных, построены распределения значений характеристик смерчеобразующих ячеек и их динамики, выделены наиболее часто встречаемые значения, удалены выбросы. Проведен анализ корреляции характеристик между собой.

С использованием полученных результатов построены модели (на базе разных методов интеллектуального анализа), классифицирующие ячейки по степени опасности возникновения из них смерчей.

При решении задачи классификации смерчевых ячеек оптимальной оказалась модель линейной классификации. Наиболее точные результаты получались на выборке, избавленной от корреляций методом удаления. Лучшее отношение безопасных ячеек к опасным: 95/100. Точность логистической регрессии и метрического классификатора не сильно уступает модели линейной классификации, поэтому можно считать эти модели также подходящими для данной задачи. Худшие результаты продемонтрировала модель решающих деревьев, поэтому без дальнейшей доработки она не может быть применена на практике.

ЛИТЕРАТУРА

[1] Wang Y., Coning E., Jacobs W., Joe P., Nikitina L., Roberts R., Wang J., Wilson J. Guidelines for nowcasting techniques. *World Meteorological Organization*, 2017, no. 1198, 82 p.

[2] Киктев Д.Б., Муравьев А.В., Смирнов А.В. Наукастинг метеорологических параметров и опасных явлений: опыт реализации и перспективы развития. *Гидрометеорологические исследования и прогнозы*, 2019, № 4, с. 92–111.

[3] Мазуров Г.И., Васильев В.А., Акселевич В.И. Анализ характеристик смерчей в России за полтора столетия. *Метеоспектр*, 2011, № 4, с. 149–155.

[4] Гмурман В.Е. *Теория вероятностей и математическая статистика: учебное пособие для вузов*. Москва, Высшая школа, 2004, 479 с.

[5] Лаврик С.А. Результаты анализа эффективности и применимости статистических методов для определения информативного набора сейсмических атрибутов. *Технологии сейсморазведки*, 2009, № 1, с. 36–44.

[6] Dimitrienko Yu.I., Koryakov M.N., Zakharov A.A. Computational modeling of conjugated aerodynamic and thermomechanical processes in composite structures of high–speed aircraft. *Applied Mathematical Sciences*, 2015, vol. 9, no. 98, pp. 4873–4880.

[7] Димитриенко Ю.И., Леонтьева С.В. Моделирование термоконвективных процессов при однонаправленной кристаллизации сплавов с учетом движения свободных границ. *Математическое моделирование и численные методы*, 2018, № 4, с. 3–24.

[8] Dimitrienko Y.I., Koryakov M.N., Zakharov A.A. Application of finite difference TVD methods in hypersonic aerodynamics. *Lecture Notes in Computer Science*, 2015, vol. 9045, pp. 161–168.

[9] Димитриенко Ю.И., Шугуан Ли Конечно–элементное моделирование неизотермического стационарного течения неньютоновской жидкости в сложных областях. *Математическое моделирование и численные методы*, 2018, № 2, с. 70– 95.

[10] Parkhomenko V.P. Modeling of global and regional climate response to solar radiation management. *Journal of Physics: Conference Series*, 2018, vol. 1141, art no. 012057. DOI: 10.1088/1742-6596/1141/1/012057
[11] Воронцов К.М. Введение в машинное обучение. *Coursera* [Электронный ресурс]. URL: https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie (дата обращения: 14.05.2021)
[12] Флах П. *Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных*. Москва, ДМК-Пресс, 2015, 400 с.
[13] Коэльо Л.П., Ричарт В. *Построение систем машинного обучения на языке Python*. Москва, ДМК Пресс, 2016, 302 с.
[14] Маккини У. *Python и анализ данных*. Москва, ДМК Пресс, 2020, 540 с.
[15] Калмыкова О.В. *Оценка смерчеопасности вблизи Черноморского побережья Краснодарского края и Республики Крым*. Дисс. канд. физ.-мат. наук. Обнинск, 2019, 230 с.
[16] Дмитриева Т.Г., Песков Б.Е. Синоптические условия, наукастинг и модельные прогнозы сильных шквалов и смерчей в Башкирии 1 июня 2007 г. и 29 августа 2014 г. *Метеорология и гидрология*, 2016, № 10, с. 16–29.
[17] Калмыкова О.В., Шершаков В.М. Индекс смерчеопасности российской акватории Черного моря. *Труды Главной геофизической обсерватории им. А.И. Воейкова*, 2017, № 584, с. 142–163.
[18] Калмыкова О.В., Шершаков В.М. Технология мониторинга смерчеопасных ситуаций на российской акватории Черного моря. *Метеорология и гидрология*, 2016, № 10, с. 93–102.
[19] Калмыкова О.В., Шершаков В.М. Технология оценки и прогноза смерчеопасности на российской акватории Черного моря и результаты ее тестирования в сезон смерчей 2017 года. *Гидрометеорологические исследования и прогнозы*, 2018, № 1 (367), с. 146–167.

Статья поступила в редакцию 15.06.2021

Ссылку на эту статью просим оформлять следующим образом:
Шершакова А.О., Пархоменко В.П. Методы интеллектуального анализа данных в модели наукастинга опасных явлений. *Математическое моделирование и численные методы*, 2021, № 3, с. 88–104.

Шершакова Анна Олеговна — студент кафедры «Вычислительная математика и математическая физика» МГТУ им. Н.Э. Баумана. e-mail: anna.shershakova@gmail.com

Пархоменко Валерий Павлович — канд. физ.-мат. наук, ведущий научный сотрудник Вычислительного центра им. А.А. Дородницына РАН Федерального исследовательского центра «Информатика и управление» РАН, доцент кафедры «Вычислительная математика и математическая физика» МГТУ им. Н.Э. Баумана. e-mail: vparhom@yandex.ru

Methods of data mining in the nowcasting model of dangerous phenomena

© A.O. Shershakova[1], V.P. Parkhomenko[1,2]

[1]Bauman Moscow State Technical University, Moscow, 105005, Russia
[2]Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS, Moscow, 119333, Russia

This work is devoted to the study and application of methods of intellectual analysis for the implementation of the scheme of the nowcasting of dangerous phenomena. In the course of the work, data sets were formed with differ in the methods of information processing for their preparation. For each set, a number of mathematical models were constructed for classifying cloud cells according to the degree of danger of tornadoes forming from them. The Python programming language has been chosen as the main development language. The work is of great practical importance in the field of forecasting weather events. Its novelty lies in the use of modern machine learning methodology, instead of the traditional approach to data extrapolation, widely used in various schemes of nowcasting.

Keywords: *data science, data mining, machine learning, mathematical model, Python, nowcasting, tornadoes, cloud cells*

REFERENCES

[1] Wang Y., Coning E., Jacobs W., Joe P., Nikitina L., Roberts R., Wang J., Wilson J. Guidelines for nowcasting techniques. *World Meteorological Organization*, 2017, no. 1198, 82 p.

[2] Kiktev D.B., Muravev A.V., Smirnov A.V. Nowcasting of meteorological parameters and hazards: implementation experience and development prospects. *Hydrometeorological Research and Forecasting*, 2019, no. 4, pp. 92–111.

[3] Mazurov G.I., Vasiliev V.A., Akselevich V.I. Analiz harakteristik smer-chej v Rossii za poltora stoletiya [Analysis of characteristics of tornadoes in Russia for a century and a half]. *Meteospektr* [Meteospectrum], 2011, no. 4, pp. 149–155.

[4] Gmurman V.E. *Teoriya veroyatnostej i matematicheskaya statistika: uchebnoe posobie dlya vuzov* [Probability theory and mathematical statistics: a textbook for universities]. Moscow, Vysshaya shkola Publ., 2004, 479 p.

[5] Lavrik S.A. Rezul'taty analiza effektivnosti i primenimosti statisticheskih metodov dlya opredeleniya informativnogo nabora sejsmicheskih atributov [Results of the analysis of the effectiveness and applicability of statistical methods for determining an informative set of seismic attributes]. *Seismic Technologies*, 2009, no. 1, pp. 36–44.

[6] Dimitrienko Yu.I., Koryakov M.N., Zakharov A.A. Computational modeling of conjugated aerodynamic and thermomechanical processes in composite structures of high–speed aircraft. *Applied Mathematical Sciences*, 2015, vol. 9, no. 98, pp. 4873–4880.

[7] Dimitrienko Y.I., Leontieva S.V. Modeling of thermal convection processes under unidirectional crystallization of alloys with liquid bridges motion. *Mathematical Modeling and Computational Methods*, 2018, no. 4, pp. 3–24.

[8] Dimitrienko Y.I., Koryakov M.N., Zakharov A.A. Application of finite difference TVD methods in hypersonic aerodynamics. *Lecture Notes in Computer Science*, 2015, vol. 9045, pp. 161–168.

[9] Dimitrienko Y.I., Li S. Mathematical simulation of non-isothermal steady flow of non-Newtonian fluid by finite element method. *Mathematical Modeling and Computational Methods*, 2018, no. 2, pp. 70–95.

[10] Parkhomenko V.P. Modeling of global and regional climate response to solar radiation management. *Journal of Physics: Conference Series*, 2018, vol. 1141, art no. 012057. DOI: 10.1088/1742-6596/1141/1/012057

[11] Vorontsov K.M. [Introduction to machine learning]. *Coursera* [Electronic resource]. URL: https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie (accessed:14.05.2021)

[12] Flach P. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press, 2012, 409 p.

[13] Coelho L.P., Richert W. *Building machine learning systems with Python*. Birmingham, Packt Publ., 2013, 290 p.
[14] McKinney W. *Python for data analysis*. USA, O'Reilly Media, Inc., 2012, 470 p.
[15] Kalmykova O.V. *Ocenka smercheopasnosti vblizi Chernomorskogo poberezh'ya Krasnodarskogo kraya i Respubliki Krym* [Assessment of tornado hazard near the Black Sea coast of the Krasnodar Territory and the Republic of Crimea]. Diss. Cand. Sc. (Phys.-Math.), Obninsk, 2019, 230 p.
[16] Dmitrieva T.G., Peskov B.E. Synoptic conditions, nowcasting, and numerical prediction of severe squalls and tornados in Bashkortostan on june 1, 2007 and august 29, 2014. *Russian Meteorology and Hydrology*, 2016, vol. 41, no. 10, pp. 673–682.
[17] Kalmykova O.V., Shershakov V.M. Waterspout risk index over the russian Black Sea water area. *Trudy Glavnoj geofizicheskoj observatorii im. A.I. Voejkova* [Proceedings of the Main Geophysical Observatory named after A.I. Voeykov], 2017, no. 584, pp. 142–163.
[18] Kalmykova O.V., Shershakov V.M. A technology of waterspout monitoring over the Russian part of the Black Sea. *Russian Meteorology and Hydrology*, 2016, vol. 41, no. 10, pp. 728–734.
[19] Kalmykova O.V., Shershakov V.M. Technology of estimation and forecasting of the risk of the waterspouts occurrence over the Russian part of the Black Sea and results of its testing during waterspouts season 2017. *Hydrometeorologica Research and Forecasting*, 2018, no. 1 (367), pp. 146–167.

Shershakova A.O., Student of Department of Computational Mathematics and Mathematical Physics, Bauman Moscow State Technical University. e-mail: anna.shershakova@gmail.com

Parkhomenko V.P., Cand. Sc. (Phys.-Math.), Leading Scientific Reaseacher, Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS, Assoc. professor of the Computational Mathematics and Mathematical Physics Department, Bauman Moscow State Technical University. e-mail: vparhom@yandex.ru