

Модифицированный непараметрический алгоритм автоматической классификации статистических данных большого объема и его применение

В. П. Тубольцев¹, А.В. Лапко^{1,2}, В.А. Лапко^{1,2}

¹ Сибирский государственный университет науки и технологий им. акад. М. Ф. Решетнева, Красноярск, Россия

² Институт вычислительного моделирования СО РАН, Красноярск, Россия

Аннотация. Предлагается модифицированный непараметрический алгоритм автоматической классификации статистических данных большого объема. По этим данным осуществляется синтез регрессионной оценки плотности вероятности. Полученная информация является основой алгоритмизации процедуры автоматической классификации. Вычислительная эффективность модифицированного непараметрического алгоритма автоматической классификации статистических данных большого объема обеспечивается процедурой сжатия исходных данных, совершенствованием и алгоритмизацией традиционной непараметрической методики обнаружения классов. Эффективность разработанной методики автоматической классификации подтверждается результатами ее применения при анализе данных дистанционного зондирования лесных массивов, поврежденных сибирским шелкопрядом.

Ключевые слова: непараметрический алгоритм автоматической классификации, регрессионная оценка плотности вероятности, дискретизация области значений случайных величин, лесные массивы, данные дистанционного зондирования.

DOI 10.14357/20718594230405

EDN QHNFURU

Введение

Методы автоматической классификации используются для обнаружения компактных групп наблюдений в статистических данных исследования объектов различной природы. Систематизация этих методов представлена в работах [1, 2]. Особое внимание уделяется разработке алгоритмов автоматической классификации, направленных на обнаружение компактных групп наблюдений (классов), соответствующих одномодальным фрагментам плотности вероятности признаков исследуемых объектов. Подобное определение класса было введено академиком Я.З. Цыпкиным [3] и развито в работах профессора В.И. Васильева [4] с использованием непараметрической оценки плотности вероятности типа Розенблатта – Парзена [5, 6].

Непараметрические системы автоматической классификации, основанные на ядерных оценках плотности вероятности, являются одним из активно развивающихся направлений теории классификации. Их практическая значимость состоит в возможности создания информационных средств, адаптируемых к условиям функционирования объектов различной природы. В работе [7] обоснована возможность решения задачи автоматической классификации с использованием алгоритма распознавания образов путем последовательного непараметрического оценивания байесовского уравнения разделяющей поверхности между классами. Класс соответствует одномодальному фрагменту совместной плотности вероятности распределения признаков классифицируемых объектов. Предложенный подход развит при решении задачи автоматической классификации в условиях

больших объемов статистических данных [8]. Его идея состоит в сжатии исходной информации путем декомпозиции пространства признаков в массив данных, составленный из центров многомерных интервалов дискретизации и соответствующих им частот принадлежности значений многомерной случайной величины.

Система дистанционного зондирования широко применяется при решении задач экологического мониторинга на основе использования алгоритмов автоматической классификации [8, 9]. Исходные статистические данные дистанционного зондирования характеризуются большими объемами спектральных признаков элементов исследуемых природных объектов.

Цель статьи состоит в повышении вычислительной эффективности непараметрического алгоритма обнаружения компактных групп наблюдений, соответствующих одномерным фрагментам плотности вероятности, на основе сжатия исходной информации и его применении при структурном анализе данных дистанционного зондирования лесных массивов, поврежденных сибирским шелкопрядом.

1. Синтез регрессионной оценки плотности вероятности

Вычислительная эффективность непараметрических алгоритмов ядерного типа во многом зависит от объема n статистических данных $V = (x_v^i, v = \overline{1, k}, i = \overline{1, n})$ размерности k и снижается по мере их увеличения. Поэтому возникает задача уменьшения объема исходных данных, которую будем решать на основе их сжатия с использованием регрессионной оценки плотности вероятности.

Методика формирования регрессионной оценки плотности вероятности предполагает выполнение следующих действий [10].

1. Разбить область определения плотности вероятности $p(x)$ многомерной случайной величины $x = (x_v, v = \overline{1, k})$ размерности k по каждой ее компоненте x_v на N_v непересекающихся интервалов длиной $2\beta_v, v = \overline{1, k}$. Для выбора количества N интервалов дискретизации используем результаты исследований [11].

2. Определить оценки вероятностей попадания элементов выборки V в каждый j многомерный интервал:

$$\bar{P}^j = n_j / n, j = \overline{1, \bar{N}},$$

где $\bar{N} = \prod_{v=1}^k N_v$; n_j - количество наблюдений

$x^j \in D_j = [z_v^j \pm \beta_v, v = \overline{1, k}]$; $z^j = (z_v^j, v = \overline{1, k})$ - координаты центра j -го интервала.

3. Будем полагать, что в каждом многомерном интервале $D_j, j = \overline{1, \bar{N}}$ случайная величина $x = (x_v, v = \overline{1, k})$ имеет равномерный закон распределения. В этих условиях вычислить оценки плотностей вероятностей:

$$\bar{p}^j(x) = \bar{P}^j = \bar{P}^j / \prod_{v=1}^k (2\beta_v) \quad \forall x \in D_j, j = \overline{1, \bar{N}}.$$

4. На основе полученной информации сформировать массив данных

$$V' = (z^j, \bar{p}^j, j = \overline{1, \bar{N}})$$

и осуществить синтез регрессионной оценки плотности вероятности $p(x)$ [10]

$$\bar{p}(x) = \frac{1}{\prod_{v=1}^k c_v} \sum_{j=1}^{\bar{N}} \bar{P}^j \prod_{v=1}^k \Phi\left(\frac{x_v - z_v^j}{c_v}\right), \quad (1)$$

где ядерные функции $\Phi(u_v)$ удовлетворяют условиям:

$$\begin{aligned} \Phi(u_v) &= \Phi(-u_v), 0 \leq \Phi(u_v) < \infty, \int \Phi(u_v) du_v = 1, \\ \int u_v^2 \Phi(u_v) du_v &= 1, \int u_v^m \Phi(u_v) du_v < \infty, 0 \leq m < \infty, \\ &v = \overline{1, k}. \end{aligned}$$

Здесь и далее бесконечные пределы интегрирования опускаются. Коэффициенты размытости c_v ядерных функций $\Phi(u_v)$ убывают с ростом N_v . Значения c_v зависят от длины интервала изменения случайных величин $x_v, v = \overline{1, k}$. Поэтому будем полагать, что $c_v = c \bar{\sigma}_v$, где $\bar{\sigma}_v$ - оценка среднего квадратического отклонения случайной величины $x_v, v = \overline{1, k}$.

Выбор оптимального значения $\bar{c}$ параметра c будем определять из условия минимума средней квадратической ошибки аппроксимации регрессионной оценкой (1) плотности вероятности $p(x)$:

$$W(c) = \frac{1}{\bar{N}} \sum_{j=1}^{\bar{N}} (\bar{P}^j - \bar{p}(z^j))^2.$$

Предлагаемая процедура оптимизации регрессионной оценки плотности вероятности случайной величины повышает вычислительную эффективность выбора коэффициентов размытости ядерных функций по сравнению с

традиционной методикой. Эта методика основана на минимизации оценки среднеквадратического критерия [11-14], который для регрессионной оценки плотности вероятности имеет вид:

$$\bar{W}(c) = \int \dots \int \bar{p}^2(x_1, \dots, x_k) dx_1 \dots dx_k - 2 \sum_{i=1}^{\bar{N}} \bar{P}^i \bar{p}(x_1^i, \dots, x_k^i).$$

Использование критерия $\bar{W}(c)$ при выборе коэффициента размытости статистики (1) приведет к значительному повышению временных затрат.

Из условия минимума среднего квадратического отклонения $\bar{p}(x)$ от $p(x)$ в работе [11] определено оптимальное количество интервалов дискретизации области значений одномерной случайной величины, соответствующее целому значению выражения:

$$N_v = \left(\Delta_v \|p(x_v)\|^2 n \right)^{1/2}, \quad v = \overline{1, k}, \quad (2)$$

где $\|p(x_v)\|^2 = \int p^2(x_v) dx_v$; Δ_v – длина интервала значений x_v .

Значение N_v определяется видом плотности вероятности $p(x_v)$ и не зависит от её параметров. Полученная закономерность является объективной, так как не зависит от вида ядерных функций оценки плотности вероятности (1). Для равномерного закона распределения одномерной случайной величины количество интервалов дискретизации совпадает с формулой Хайнхольда-Гаеде [15]: $N = \sqrt{n}$.

Результаты оценивания интеграла от квадрата плотности вероятности случайной величины в выражении (2) по коэффициенту контрэксцесса представлены в работе [11].

Регрессионная оценка плотности вероятности является основой синтеза непараметрического алгоритма автоматической классификации в условиях большого объема статистических данных.

2. Непараметрический алгоритм автоматической классификации

Пусть имеется выборка многомерных статистических данных $V = (x_v^i, v = \overline{1, k}, i = \overline{1, n})$ большого объема n , распределенная с неизвестной плотностью вероятности $p(x) = p(x_1, \dots, x_k)$. Необ-

ходимо выборку V разбить на группы компактных наблюдений $V_j, j = \overline{1, M}$, количество M которых неизвестно. Под компактной группой наблюдений (класс) будем понимать их множество, которое соответствует одномодальному фрагменту плотности вероятности $p(x)$.

Следуя методике синтеза регрессионной оценки плотности вероятности, преобразуем исходную выборку V в массив данных $\bar{V} = (z^i, \bar{P}^i, i = \overline{1, \bar{N}})$.

Для обнаружения первого класса по информации массива данных $\bar{V}$ выберем элемент $(z^j, \bar{P}^j)$, который определяется условием:

$$\bar{P}^j = \max_{t=1, \bar{N}} \bar{P}^t. \quad (3)$$

Этот элемент $(z^j, \bar{P}^j)$ соответствует максимальной частоте встречаемости случайной величины в многомерных интервалах. Тогда начальный этап формирования элементов выборки $\bar{V}$, принадлежащих первому классу Ω_1 , определяется правилом:

$$z^i \in \Omega_1, \text{ если } \prod_{v=1}^k \Phi\left(\frac{z_v^i - z_v^j}{c_v}\right) R(\bar{P}^i, \bar{P}^j) > 0, \quad (4)$$

$$i \in I = (t = \overline{1, \bar{N}}), i \neq j.$$

При ступенчатой ядерной функции

$$\Phi\left(\frac{z_v^i - z_v^j}{c_v}\right) = \begin{cases} 0,5, & \text{если } |z_v^i - z_v^j| \leq c_v \\ 0, & \text{если } |z_v^i - z_v^j| > c_v. \end{cases}$$

Индикаторная функция в выражении (4) принимает значения:

$$R(\bar{P}^i, \bar{P}^j) = \begin{cases} 1, & \text{если } \bar{P}^i \leq \bar{P}^j \\ 0, & \text{если } \bar{P}^i > \bar{P}^j. \end{cases}$$

При значениях $c_v \geq 2\beta_v$ определяются элементы массива $\bar{V}$ смежные с элементом $(z^j, \bar{P}^j)$. Индикаторная функция $R(\bar{P}^i, \bar{P}^j)$ позволяет определять элементы $(z^i, \bar{P}^i)$ массива данных $\bar{V}$, характеризуется убывающими значениями плотности вероятности x в окрестности z^j . На этом этапе значения $R(\bar{P}^i, \bar{P}^j) = 1$, так как выполняется условие (3).

Обозначим через I_1 множество номеров элементов массива данных $\bar{V}$, отнесенных на первом этапе автоматической классификации в соответствии с правилом (4) к первому классу Ω_1 . Множество I_1 включает номер j элемента $(z^j, \bar{P}^j)$.

Второй этап классификации осуществляется решающим правилом:

$$\begin{aligned} & z^i \in \Omega_1, \text{ если} \\ & \sum_{j \in I_1} \prod_{v=1}^k \Phi \left(\frac{z_v^i - z_v^j}{c_v} \right) R(\bar{P}^i, \bar{P}^j) > 0, \quad (5) \\ & i \in I \setminus I_1. \end{aligned}$$

По аналогии на t -ом этапе автоматической классификации элементы выборки $\bar{V}$, принадлежащие первому классу Ω_1 , определяются правилом:

$$\begin{aligned} & z^i \in \Omega_1, \text{ если} \\ & \sum_{j \in I_{t-1}} \prod_{v=1}^k \Phi \left(\frac{z_v^i - z_v^j}{c_v} \right) R(\bar{P}^i, \bar{P}^j) > 0, \quad (6) \\ & i \in I \setminus (I_1 \cup I_2 \dots \cup I_{t-1}). \end{aligned}$$

Последовательная процедура автоматической классификации (4)-(6) продолжается до выполнения условия $I_t = \emptyset$, когда отсутствует обнаружение новых элементов массива данных $\bar{V}$ из класса Ω_1 . Обнаруженный первый класс характеризуется регрессионной оценкой плотностью вероятности

$$\bar{p}_1(x) = \frac{1}{\prod_{v=1}^k c_v} \sum_{i \in I_{t-1}} \bar{P}^i \prod_{v=1}^k \Phi \left(\frac{x_v - z_v^i}{c_v} \right),$$

для восстановления которой используются данные: $\bar{V}_1 = (z^i, \bar{P}^i, i \in I_{t-1})$.

Для обнаружения элементов массива $\bar{V}_2 \in \bar{V}$, принадлежащих второму классу Ω_2 , из оставшихся данных $\bar{V} \setminus \bar{V}_1$ выбирается элемент $(z^j, \bar{P}^j)$ с максимальным значением $\bar{P}^j$ и предложенная выше процедура автоматической классификации используется для определения класса Ω_2 . По аналогии формируются оставшиеся классы, количество которых априори не определено.

После завершения процесса автоматической классификации оставшиеся элементы массива данных $\bar{V}$ и классы с минимальным количеством элементов требуют дополнительного анализа для

обоснования причин их появления. Например, такими причинами могут являться появление источников начального поражения лесных массивов шелкопрядом, наличие вырубок древостоев либо их усыхания.

3. Оценивание состояний лесных массивов по данным дистанционного зондирования

Территория исследования определялась восточной частью Ибейского района Красноярского края. На этой территории преобладают пихтовые, кедровые и еловые древостои, встречаются березы и осины. Высоты варьируются в диапазоне от 500 до 900 метров над уровнем моря. В период массового размножения сибирским шелкопрядом на этой территории в 2019-2020 гг. было повреждено более 1500 гектаров древостоев. Часть поврежденных лесных массивов располагается в Иркутской области.

Исходные данные для обработки представляют собой фрагмент мультиспектрального изображения площадью 13,8 тысяч гектар (Рис. 1, а). Изображение собрано из шести спектральных каналов съемки аппарата Европейского космического агентства Sentinel-2A за 1 июня 2022 года. Архив с данными дистанционного зондирования Земли получен с геопортала Earth Explorer. Каждый пиксель исходных данных характеризуется шестью спектральными признаками $x = (x_1, \dots, x_6)$, которым соответствуют длины волн (нанометры): 492,7 (x_1), 559,8 (x_2), 664,6 (x_3), 740,5 (x_4), 832,8 (x_5), 2202,4 (x_6), а общее их количество равно 1286776. Данные дистанционного зондирования обработаны в свободной кроссплатформенной геоинформационной системе QGIS при помощи бесплатного плагина с открытым исходным кодом Semi-Automatic Classification Plugin, реализующего алгоритм атмосферной коррекции Dark Object Subtraction (DOS) [16]. На Рис. 1, а представлено RGB-изображение исходных данных, где каналам R, G, B соответствуют признаки x_6, x_5, x_3 .

Для обнаружения компактных групп наблюдений (классов) в пространстве спектральных признаков $x = (x_1, \dots, x_6)$ использовался предлагаемый непараметрический алгоритм автоматической классификации. На его основе осуществлена корректировка программы автоматической классификации NAC v.2.0 [17], которая позволяет

загружать изображения в формате GeoTIFF, производить классификацию и получать изображение для дальнейшей геообработки. В этой программе реализованы методы многопоточных вычислений, что позволило повысить скорость обработки данных дистанционного зондирования. Обработка данных проводилась на персональном компьютере со следующими характеристиками: процессор Intel(R) Core(TM) i5-10500, объем оперативной памяти 16 ГБ, жесткий диск Seagate Exos 7E8. Время решения

задачи автоматической классификации предлагаемым непараметрическим алгоритмом автоматической классификации (программа NAC v.2.0) составило 29 секунд.

Развитие функциональных возможностей программы NAC v.2.0. позволило в исходном изображении обнаружить 320 компактных групп точек (Рис. 1, в). Из них 275 классов имеет объем менее 0,1% от количества пикселей исходного изображения, а 45 классов содержали более 12868 пикселей. Обнаруженные классы

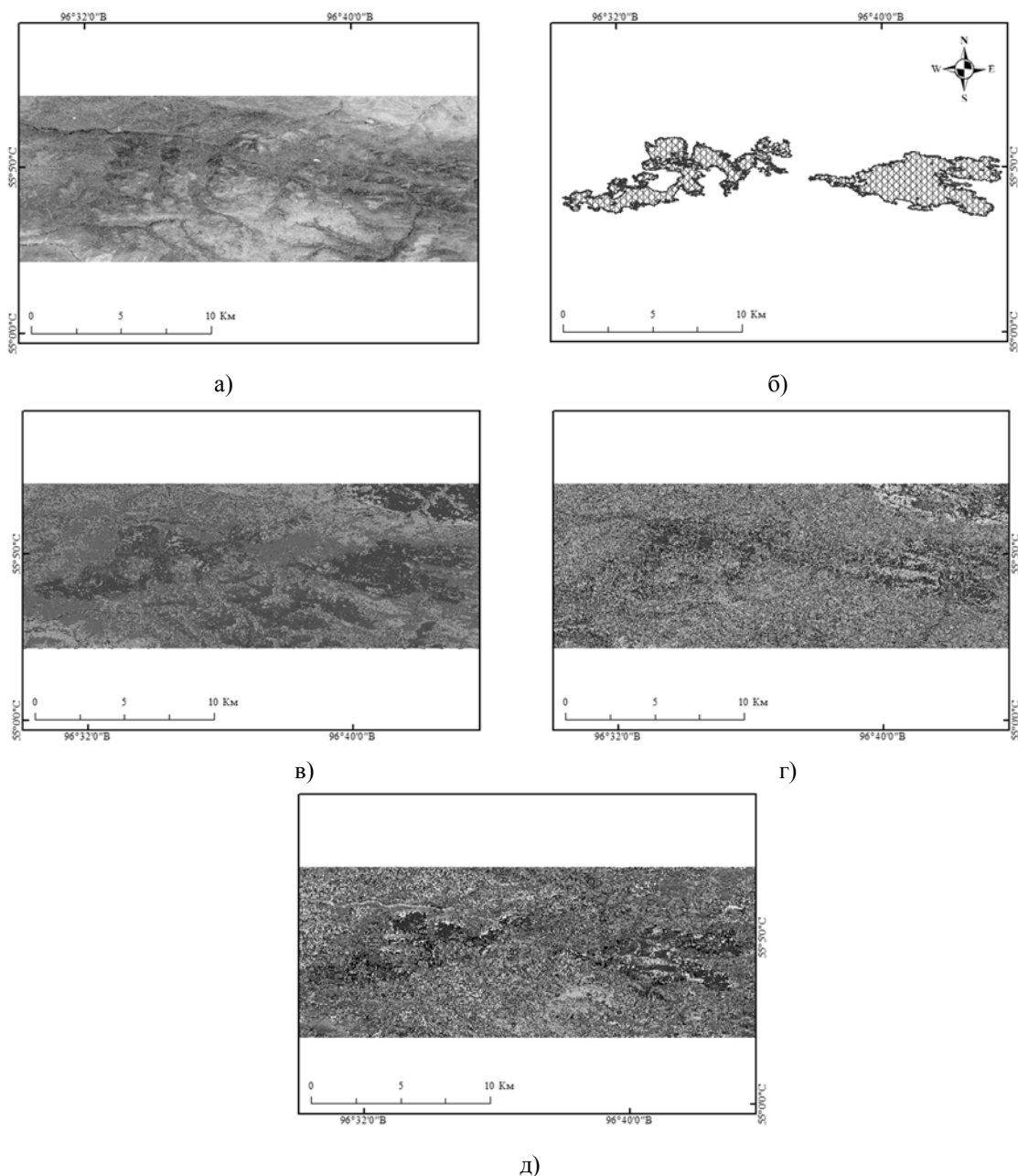


Рис. 1. Сопоставление исходного снимка (а) с результатами классификации непараметрическим алгоритмом (в), методом ISODATA при 237 классах (г), и при 45 классах (д). Отображение векторного представления поврежденных лесных насаждений сибирским шелкопрядом (б)

соответствуют лесным массивам с различной степенью поражения сибирским шелкопрядом, усохшим и листовным древостоям, травяно-кустарниковым сообществам и заболоченным участкам исследуемой территории.

Результаты автоматической классификации в указанных условиях методом ISODATA средствами программного пакета ArcGIS ArcMap [18] приведены на Рис. 1, г, д. При настройке этого инструмента было указано количество классов, равное обнаруженным непараметрическим алгоритмом автоматической классификации. В результате было получено классифицированное изображение с 237 компактными группами точек (Рис. 1, г). Пиксели между ними были распределены неравномерно, но минимальное их количество в классах было не менее 0,1% от общего объема исходных данных. На Рис. 1, д приставлены результаты работы алгоритма ISODATA при указанном количестве классов равным 45.

В центральной части исследуемой территории находятся лесные массивы усохших темнохвойных древостоев и поврежденных сибирским шелкопрядом. Они отображаются на рисунке ярким розовым цветом. Фрагменты рисунка зеленого цвета, от темных к светлым тонам, соответствуют светлохвойным, листовным древостоям и травяно-кустарниковым сообществам. Синий оттенок имеют водные объекты, бирюзовый цвет соответствует заболоченным участкам. Экспертный анализ показал, что предлагаемый метод автоматической классификации обладает преимуществом по сравнению с алгоритмом ISODATA, что следует из анализа рисунков. Сравнение исходного снимка с результатами автоматической классификации осуществлялся специалистами Центра защиты леса Красноярского края. Исходный снимок (Рис. 1, а) более близок к Рис. 1, в, чем к Рис. 1, г. Рассматриваемые алгоритмы автоматической классификации достаточно полно выделяют зоны поврежденных древостоев, разделяют темнохвойные и светлохвойные породы, определяют заболоченные территории и участки с открытой почвой. Алгоритм ISODATA в этих условиях показал более детальный результат классификации, чем предлагаемый непараметрический алгоритм автоматической классификации. Это можно объяснить тем, что последний основан на определении класса как одномодального фрагмента плотности вероятности

спектральных признаков исследуемого объекта, аппроксимационные свойства которой зависят от количества многомерных интервалов дискретизации. По результатам автоматической классификации алгоритмом ISODATA и непараметрическим классификатором экспертами были определены площади территорий поврежденных древостоев сибирским шелкопрядом (Рис. 1, б). Они составили 1825 и 1934,5 гектаров лесных массивов соответственно, что указывает на их различие в 6%.

Заключение

Модифицированный непараметрический алгоритм автоматической классификации статистических данных большого объема основан на их сжатии путем декомпозиции многомерного пространства спектральных признаков исследуемых объектов, совершенствовании и алгоритмизации традиционной непараметрической процедуры автоматической классификации. Предлагаемая методика позволяет преобразовать исходную статистическую информацию в массив данных, составленный их центров многомерных интервалов дискретизации и соответствующих им частот принадлежности случайных величин. Основу модифицированного алгоритма автоматической классификации составляет формализация процедуры проверки близости центров многомерных интервалов дискретизации и соотношений между частотами их встречаемости. Его применение позволяет обнаружить классы, соответствующие одномодальным фрагментам плотности вероятности многомерной случайной величины. Вычислительная эффективность предлагаемого алгоритма автоматической классификации обосновывается формализацией процедуры обнаружения компактных групп наблюдений и новой методикой оптимизации регрессионной оценки плотности вероятности.

Предлагаемый метод позволяет сократить время классификации при использовании многопоточной технологии обработки данных, что подтверждается результатами обработки спектральных данных дистанционного зондирования лесных массивов, поврежденных сибирским шелкопрядом. Результаты автоматической классификации показывают преимущество непараметрического алгоритма по сравнению с методом программного продукта ArcGIS ArcMap.

Литература

1. Дорофеев А.А. Алгоритмы автоматической классификации (обзор) // Автоматика и телемеханика. 1971. № 12. С. 78-113.
2. Дорофеев А.А. Методология экспертно-классификационного анализа в задачах управления и обработки сложноорганизованных данных (история и перспективы развития) // Проблемы управления. 2009. № 3.1. С. 19-28.
3. Цыпкин Я.З. Основы теории обучающихся систем. М.: Наука, 1970.
4. Васильев В.И., Эш С.Н. Особенности алгоритмов самообучения и кластеризации // Управляющие системы и машины. 2011. № 3. С. 3-9.
5. Parzen E. On estimation of a probability density function and mode // Annals of Mathematical Statistics. 1962. V. 33. No 3. P. 1065-1076.
6. Епанечников В.А. Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. Т. 14. № 1. С. 156-161.
7. Лапко А.В., Лапко В.А. Непараметрический алгоритм автоматической классификации в условиях статистических данных большого объема // Информатика и системы управления. 2018. Т. 57. № 3. С. 59-70.
8. Зеньков И.В., Лапко А.В., Лапко В.А., Им С.Т., Тубольцев В.П., Авдеев В.Л. Непараметрический алгоритм автоматической классификации многомерных статистических данных большого объема и его применение // Компьютерная оптика. 2021. Т. 45. № 2. С. 253-260.
9. Васильева И.К., Попов А.В. Метод автоматической кластеризации данных дистанционного зондирования // Авиационно-космическая техника и технология. 2019. Т. 155. № 3. С. 64-75.
10. Лапко А.В., Лапко В.А. Ядерные оценки плотности вероятности и их применение. Красноярск: СибГУ им. М.Ф. Решетнёва. 2021.
11. Rudemo M. Empirical choice of histogram and kernel density estimators // Scandinavian Journal of Statistics. 1982. V. 9. No 2. P. 65-78.
12. Hall P. Large-sample optimality of least squares cross-validation in density estimation // Annals of Statistics. 1983. V. 11. No 4. P. 1156-1174.
13. Bowman A.W. An alternative method of cross-validation for the smoothing of density estimates // Biometrika. 1984. V. 71. No 2. P. 353-360.
14. Heinhold I., Gaede K.W. Ingenieur statistic. München: Wien, Springer Verlag, 1964.
15. Chavez P.S. Image-based atmospheric correction - revisited and improved // Photogrammetric Engineering and Remote Sensing. 1996. V. 62. No 9. P. 1025-1036.
16. Лапко А.В., Лапко В.А., Им С.Т., Тубольцев В.П., Авдеев В.Л. Программа автоматической классификации данных дистанционного зондирования Земли на основе непараметрических алгоритмов принятия решений (НАС v. 2.0) // Свидетельство о государственной регистрации программы для ЭВМ № 2022619023 от 18.05.2022. Федеральная служба по интеллектуальной собственности, 2022.
17. Lemenkova P. ISO Cluster classifier by ArcGIS for unsupervised classification of the Landsat TM image of Reykjavík. University thought // Bulletin of Natural Sciences Research. 2021. V. 11. No 1. P. 29-37.

Тубольцев Виталий Павлович. Аспирант. Сибирский государственный университет науки и технологий им. акад. М.Ф. Решетнёва. Области исследований: разработка информационных систем, непараметрические системы классификации, быстрые алгоритмы оптимизации решающих правил, обработка данных дистанционного зондирования. E-mail: vitalitya.98@mail.ru

Лапко Александр Васильевич. Доктор технических наук, профессор. Главный научный сотрудник. Институт вычислительного моделирования Сибирского отделения Российской академии наук. Области исследований: непараметрическая статистика, распознавания образов и анализ изображений, моделирование и оптимизация неопределенных систем, дистанционное зондирование. E-mail: lapko@icm.krasn.ru (Ответственный за переписку)

Лапко Василий Александрович. Доктор технических наук, профессор. Ведущий научный сотрудник. Институт вычислительного моделирования Сибирского отделения Российской академии наук. Области исследований: непараметрическая статистика, распознавания образов и анализ изображений, моделирование неопределенных систем, дистанционное зондирование. E-mail: valapko@yandex.ru

Modified Nonparametric Algorithm for Automatic Classification of Large-Volume Statistical Data and its Application

V. P. Tuboltsev¹, A. V. Lapko^{1,2}, V. A. Lapko^{1,2}

¹ Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia

² Institute of Computational Modelling SB RAS, Krasnoyarsk, Russia

Abstract. A modified nonparametric algorithm for automatic classification of large-volume statistical data is proposed. Its application makes it possible to detect classes corresponding to unimodal fragments of the probability density of a multidimensional random variable. The compression of the initial information is carried out on the basis of the decomposition of the multidimensional space of features into a

data array composed of the centers of the sampling intervals and the corresponding frequencies of belonging to the values of the random variable. Based on these data, a regression estimate of the probability density is synthesized. The information obtained is the basis for the algorithmization of the automatic classification procedure. A class is a compact group of observations of a random variable corresponding to a single-modal fragment of probability density. The computational efficiency of the modified non-parametric algorithm for automatic classification of large-volume statistical data is provided by the compression procedure of the source data, improvement and algorithmization of the traditional nonparametric method of class detection. The computational efficiency of the modified non-parametric algorithm for automatic classification of large volume statistical data is provided by the initial data compression procedure, improvement and algorithmization of the traditional nonparametric method for detecting compact groups of observations of a random variable. The effectiveness of the developed method of automatic classification is confirmed by the results of its application in the analysis of remote sensing data of forests damaged by the Siberian silkworm.

Keywords: nonparametric algorithm for automatic classification, regression estimation of probability density, discretization of the range of random variables, woodlands, remote sensing data.

DOI 10.14357/20718594230405 EDN QHNFRU

References

1. Dorofeyuk A.A. Algoritmy avtomaticheskoy klassifikatsii (obzor) [Algorithms of automatic classification (review)] // *Avtomatika i telemekhanika* [Automation and Remote Control]. 1971. No 12. P. 78-113.
2. Dorofeyuk A.A. Metodologiya ekspertno-klassifikatsionnogo analiza v zadachakh upravleniya i obrabotki slozhnoorganizovannykh dannykh (istoriya i perspektivy razvitiya) [Methodology of expert classification analysis in the management and processing of complex data (history and prospects of development)] // *Problemy upravleniya* [Control sciences]. 2009. No 3.1. P. 19-28.
3. TSypkin Ya.Z. Osnovy teorii obuchayushchikhsya sistem [Fundamentals of the theory of learning systems]. Moscow: Nauka, 1970.
4. Vasil'ev V.I., EHsh S.N. Osobennosti algoritmov samoobucheniya i klasterizatsii [Features of self-learning algorithms and clustering] // *Upravlyayushchiye sistemy i mashiny* [Control systems and machines]. 2011. No 3. P. 3-9.
5. Parzen E. On estimation of a probability density function and mode // *Annals of Mathematical Statistics*. 1962. V. 33. No 3. P. 1065-1076.
6. Epanechnikov V.A. Neparametricheskaya ocenka mnogomernoy plotnosti veroyatnosti [Non-parametric estimation of a multivariate probability density]. // *Teoriya veroyatnosti i ee primeneniya* [Theory of Probability & Its Applications]. 1969. V. 14. No 1. P. 156-161.
7. Lapko A.V., Lapko V.A. Neparametricheskii algoritm avtomaticheskoy klassifikatsii v usloviyakh statisticheskikh dannykh bol'shogo ob'yema [Nonparametric algorithm of automatic classification under conditions of large-scale statistical data] // *Informatika i sistemy upravleniya* [Informatics and control systems]. 2018. V. 57. No 3. P. 59-70.
8. Zenkov I.V., Lapko A.V., Lapko V.A., Im S.T., Tuboltsev V.P., Avdeenok V.L. A nonparametric algorithm for automatic classification of large multivariate statistical data sets and its application // *Computer Optics*. 2021. V. 45. No 2. P. 253-260.
9. Vasilyeva I.K., Popov A.V. Metod avtomaticheskoy klasterizatsii dannykh distantsionnogo zondirovaniya [Method for automatic clustering of remote sensing data] // *Aviacionno-kosmicheskaya tekhnika i tekhnologiya* [Aerospace Technic and Technology]. 2019. V. 155. No 3. P. 64-75.
10. Lapko A.V., Lapko V.A. Regressionnaya ocenka mnogomernoy plotnosti veroyatnosti i eyo svoystva [Regression estimate of the multidimensional probability density and its properties] // *Avtometriya* [Optoelectronics, Instrumentation and Data Processing]. 2014. V. 50. No 2. P.148-153.
11. Lapko A.V., Lapko V.A. Yadernyye otsenki plotnosti veroyatnosti i ikh primeneniye [Kernel probability density estimates and their application]. Krasnoyarsk: Reshetnev University. 2021.
12. Rudemo M. Empirical choice of histogram and kernel density estimators // *Scandinavian Journal of Statistics*. 1982. V. 9. No 2. P. 65-78.
13. Hall P. Large-sample optimality of least squares cross-validation in density estimation // *Annals of Statistics*. 1983. V. 11. No 4. P. 1156-1174.
14. Bowman A.W. An alternative method of cross-validation for the smoothing of density estimates // *Biometrika*. 1984. V. 71. No 2. P. 353-360.
15. Heinhold I., Gaede K.W. Ingenieur statistic. München: Wien, Springer Verlag Publs, 1964.
16. Chavez P.S. Image-based atmospheric correction - revisited and improved // *Photogrammetric Engineering and Remote Sensing*. 1996. V. 62. No 9. P. 1025-1036.
17. Lapko A.V., Lapko V.A., Im S.T., Tuboltsev V.P., Avdeenok V.L. Programma avtomaticheskoy klassifikatsii dannykh distantsionnogo zondirovaniya Zemli na osnove neparametricheskikh algoritmov prinyatiya resheniy (NAC v. 2.0) [The program for automatic classification of Earth remote sensing data based on nonparametric decision-making algorithms (NAC v. 2.0)] // Certificate of state registration of the computer program RF No. 2022619023, 2022.
18. Lemenkova P. ISO Cluster classifier by ArcGIS for unsupervised classification of the Landsat TM image of Reykjavik. University thought // *Bulletin of Natural Sciences Research*. 2021. V. 11. No 1. P. 29-37.

Tuboltsev Vitaliy. P. Graduate student. Reshetnev Siberian State University of Science and Technology. Research areas: development of information systems, nonparametric classification systems, fast algorithms for optimizing decision rules, remote sensing data processing. E-mail: vitalya.98@mail.ru

Lapko Aleksander V. Doctor of technical sciences, professor. Chief researcher. Institute of computational modelling of the Siberian branch of the Russian Academy of Sciences. Research areas: nonparametric statistics, pattern recognition and image analysis, modeling and optimization of indefinite systems, remote sensing. E-mail: lapko@icm.krasn.ru

Lapko Vasiliy A. Doctor of technical sciences, professor. Leading researcher. Institute of computational modelling of the Siberian branch of the Russian Academy of Sciences. Research areas: nonparametric statistics, pattern recognition and image analysis, modeling and optimization of indefinite systems, remote sensing. E-mail: valapko@yandex.ru