# Непараметрический алгоритм автоматической классификации многомерных статистических данных большого объёма и его применение

*И.В. Зеньков [1,5], А.В. Лапко [2,4], В.А. Лапко [2,4], С.Т. Им [1,3,4], В.П. Тубольцев [4], В.Л. Авдеенок [4]*
*[1] Сибирский федеральный университет,*
*660041, г. Красноярск, Россия, просп. Свободный, д. 79, стр. 3,*
*[2] Институт вычислительного моделирования СО РАН,*
*660036, Россия, г. Красноярск, Академгородок, д. 50, стр. 44,*
*[3] Институт леса им. В.Н. Сукачева СО РАН,*
*660036, Россия, г. Красноярск, Академгородок, д. 50, стр. 28,*
*[4] Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева,*
*660037, г. Красноярск, просп. «Красноярский рабочий», д. 31,*
*[5] Федеральный исследовательский центр информационных и вычислительных технологий,*
*660049, Россия, г. Красноярск, просп. Мира, д. 53*

***Аннотация***

Предлагается непараметрический алгоритм автоматической классификации статистических данных большого объёма. Основу алгоритма составляет процедура оптимальной дискретизации области значений случайной величины. Под классом понимается компактная группа наблюдений случайной величины, соответствующих одномодальному фрагменту плотности вероятности. Рассматриваемый алгоритм автоматической классификации основан на «сжатии» исходной информации на основе декомпозиции многомерного пространства признаков. В результате статистическая выборка большого объёма преобразуется в массив данных, составленный из центров многомерных интервалов дискретизации и соответствующих им частот принадлежности случайных величин. Для обоснования процедуры оптимальной дискретизации используются результаты исследования асимптотических свойств регрессионной оценки плотности вероятности ядерного типа. Из условия минимума среднеквадратического отклонения регрессионной оценки плотности вероятности определяются оптимальные количества интервалов дискретизации области значений одномерной и двухмерной случайных величин. Полученные результаты обобщаются на дискретизацию области значений многомерной случайной величины. Формула оптимальной дискретизации содержит составляющую, которая характеризуется нелинейным функционалом от плотности вероятности. Устанавливается аналитическая зависимость обнаруженной составляющей от коэффициента контрэксцесса одномерной случайной величины. Для независимых компонент многомерной случайной величины определяется методика расчёта оценок оптимального количества интервалов дискретизации случайных величин и их длин. На этой основе разрабатывается непараметрический алгоритм автоматической классификации, который основан на последовательной процедуре проверки близости центров многомерных интервалов дискретизации и соотношений между частотами принадлежности случайных величин из исходной выборки этим интервалам. Для дополнительного повышения вычислительной эффективности предлагаемого алгоритма автоматической классификации используется многопоточный метод его программной реализации. Практическая значимость разработанных алгоритмов подтверждается результатами их применения при обработке данных дистанционного зондирования.

*Ключевые слова*: алгоритм автоматической классификации, многомерная гистограмма, регрессионная оценка плотности вероятности, дискретизация области значений случайной величины, выборки большого объёма, коэффициент контрэксцесса, данные дистанционного зондирования.

*Цитирование*: **Зеньков, И.В.** Непараметрический алгоритм автоматической классификации многомерных статистических данных большого объёма и его применение / И.В. Зеньков, А.В. Лапко, В.А. Лапко, С.Т. Им, В.П. Тубольцев, В.Л. Авдеенок // Компьютерная оптика. – 2021. – Т. 45, № 2. – С. 253-260. – DOI: 10.18287/2412-6179-CO-801.

*Citation*: Zenkov IV, Lapko AV, Lapko VA, Im ST, Tuboltsev VP, Avdeenok VL. A nonparametric algorithm for automatic classification of large multivariate statistical data sets and its application. Computer Optics 2021; 45(2): 253-260. DOI: 10.18287/2412-6179-CO-801.

### *Введение*

Обнаружение компактных групп наблюдений в статистических данных является первоначальной задачей исследования закономерностей, свойственных объектам различной природы, которая решается алгоритмами автоматической классификации. Систематизация методов автоматической классификации представлена в работах [1, 2].

Активно развивается направление синтеза алгоритмов автоматической классификации, направленных на обнаружение компактных групп наблюдений (классов), соответствующих одномодальным фрагментам плотности вероятности признаков исследуемых объектов. Подобное определение класса было введено Я.З. Цыпкиным [3] и развито в работах В.И. Васильева с использованием непараметрической оценки плотности вероятности случайных величин [4].

В работах [5] обоснована возможность решения задачи автоматической классификации в рамках задачи распознавания образов с помощью итерационной процедуры последовательного непараметрического оценивания байесовских уравнений разделяющих поверхностей между классами, которые соответствуют одномодальным симметричным фрагментам совместной плотности вероятности распределения признаков классифицируемых объектов. Предложенный подход развит при решении задачи автоматической классификации в условиях больших объёмов статистических данных [6]. Его идея состоит в «сжатии» исходной информации путём декомпозиции пространства признаков в массив данных, состоящий из центров многомерных интервалов дискретизации и соответствующих им частот принадлежности случайных величин.

Цель данной статьи состоит в развитии непараметрических алгоритмов автоматической классификации статистических данных большого объёма для обнаружения классов, соответствующих одномодальным фрагментам плотности вероятности. Их синтез основан на использовании новой методики декомпозиции области значений многомерной случайной величины и применении технологии параллельных вычислений при разработке программных средств анализа данных дистанционного зондирования.

### *Методика дискретизации области значений многомерной случайной величины*

Для анализа законов распределения многомерных случайных величин $x=(x_v, v=\overline{1,k})$ в условиях статистических данных $V=(x^i, i=\overline{1,n})$ большого объёма $n$ используется модификация непараметрической оценки плотности вероятности $\overline{p}(x)$ [7]. Синтез $\overline{p}(x)$ основан на «сжатии» исходной информации $V$ путём декомпозиции пространства значений $(x_v, v=\overline{1,k})$ на многомерные интервалы. В результате исходная выборка $V$ преобразуется в массив данных $\overline{V}=(z^j, \overline{P}^j, j=\overline{1,N})$, составленный из центров $z^j=(z_v^j, v=\overline{1,k})$ многомерных интервалов дискретизации в количестве $N$ и соответствующих им частот $\overline{P}^j$ появления случайных величин. Полученные данные $\overline{V}$ позволяют оценить плотность вероятности $p(x)$ многомерной случайной величины $x=(x_v, v=\overline{1,k})$ в виде статистики

$$\overline{p}(x_1,\ldots,x_k)=\frac{1}{\prod\limits_{v=1}^{k}c_v}\sum_{j=1}^{N}\overline{P}^j\prod_{v=1}^{k}\Phi\left(\frac{x_v-z_v^j}{c_v}\right). \qquad (1)$$

Ядерные функции $\Phi(u_v)$ в статистике (1) удовлетворяют условиям:

$$\Phi(u_v)=\Phi(-u_v),\ 0\le\Phi(u_v)<\infty,$$
$$\int_{-\infty}^{\infty}\Phi(u_v)du_v=1,\ \int_{-\infty}^{\infty}u_v^2\,\Phi(u_v)du_v=1,$$
$$\int_{-\infty}^{\infty}u_v^m\,\Phi(u_v)du_v<\infty,\ 0\le m<\infty,$$
$$v=\overline{1,k}.$$

Коэффициенты размытости $c_v=c_v(n)$, $v=\overline{1,k}$, ядерных функций убывают с ростом объёма $n$ исходных статистических данных.

Рассматриваемая непараметрическая оценка плотности вероятности (1) относится к семейству нормированных функций, обладает свойствами асимптотической несмещённости и состоятельности.

Из условия минимума асимптотического выражения среднеквадратического отклонения $\overline{p}(x)$ от $p(x)$ по значению $N$ получены оптимальные формулы дискретизации области значений одномерной и двухмерной случайных величин [8, 9]. В частности, для одномерной случайной величины предлагаемая формула близка к формуле Гаеде и совпадает с ней при оценивании плотности вероятности с равномерным законом распределения $x$. Разработанная методика была использована при определении оптимального количества $N_k^*$ интервалов дискретизации области значений многомерной случайной величины

$$N_k^*=\left(\alpha(k)\,n\prod_{v=1}^{k}\Delta_v\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty}p^2(x_1,\ldots,x_k)\,dx_1\ldots dx_k\right)^{1/2}= \qquad (2)$$
$$=\alpha_k\sqrt{n},$$

где $\Delta_v$ – длина интервала значений $x_v$, $v=\overline{1,k}$. Коэффициент $\alpha(k)=(2k-1)/k^2\le 1$ и его значения уменьшаются с ростом размерности $k$ случайной величины.

В работе [10] вычисляются оптимальные количества интервалов дискретизации для плотностей вероятностей конкретного вида (равномерные, нормальные, экспоненциальные) и отмечается перспективность создания общей методики.

Предположим, что случайные величины $x_v$, $v=\overline{1,k}$, независимые. Тогда выражение (2) запишется в виде

$$N_k^* = \left( \prod_{v=1}^{k} \left( \left( \frac{2k-1}{k^2} n \right)^{\frac{1}{k}} \Delta_v \int_{-\infty}^{+\infty} p^2(x_v)\, dx_v \right) \right)^{1/2} = \prod_{v=1}^{k} N^*(v),$$

где

$$N^*(v) = \left( \left( \frac{2k-1}{k^2} n \right)^{\frac{1}{k}} \Delta_v \int_{-\infty}^{+\infty} p^2(x_v)\, dx_v \right)^{1/2}. \quad (3)$$

В работе [11] показано, что для одномерного случая выражение

$$W_2 = \left( \Delta_v \int_{-\infty}^{+\infty} p^2(x_v)\, dx_v \right)^{1/2}$$

может быть оценено по значению коэффициента контрэксцесса $\delta_v$ функциональными зависимостями

$$\bar{W}_2(\delta_v) = 31{,}27\delta_v^4 - 62{,}73\delta_v^3 + 46{,}34\delta_v^2 - 16{,}18\delta_v + 3{,}8, \quad (4)$$

либо

$$\bar{W}_2(\delta_v) = 1{,}06\delta_v^{-0{,}386}. \quad (5)$$

В этих условиях средняя относительная ошибка аппроксимации для модели (4) определяется значением 0,0275, а для модели (5) – 0,037. Оценивание коэффициента контрэксцесса $\delta_v$ осуществляется по каждой компоненте $x_v$, $v = \overline{1,k}$.

Тогда методика дискретизации области значений случайной величины $x = (x_v, v = \overline{1,k})$ на многомерные интервалы предполагает выполнение следующих действий:

1. По исходным статистическим данным $V = (x^i, i = \overline{1,n})$ определить оценки $\bar{\Delta}_v$ длин интервалов изменения значений случайных величин как разности между минимальными и максимальными значениями $x_v^i$, $i = \overline{1,n}$, и оценить коэффициенты контрэксцесса

$$\bar{\delta}_v = 1/\sqrt{\bar{\eta}_v},$$

где

$$\bar{\eta}_v = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_v^i - \bar{x}_v\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_v^i - \bar{x}_v\right)^2\right)^2}, \quad \bar{x}_v = \frac{1}{n}\sum_{i=1}^{n} x_v^i, \quad v = \overline{1,k}.$$

2. Используя формулу (4) либо (5), вычислить оценку константы $\bar{W}_2(\bar{\delta}_v)$, $v = \overline{1,k}$.

3. По значениям $\bar{\Delta}_v$, $\bar{W}_2(\bar{\delta}_v)$, $v = \overline{1,k}$, $n$ и $k$ в соответствии с выражением (3) определить количество интервалов дискретизации случайных величин

$$\bar{N}^*(v) = \left( \left( \frac{2k-1}{k^2} n \right)^{\frac{1}{k}} \bar{W}_2(\bar{\delta}_v) \right)^{1/2}$$

и их длину

$$\bar{\bar{\Delta}}_v = \bar{\Delta}_v / \bar{N}^*(v), \quad v = \overline{1,k}.$$

### *Непараметрический алгоритм обнаружения компактных групп наблюдений*

Имеются статистические данные $V = (x^i, i = \overline{1,n})$ наблюдений случайных величин $x = (x_v, v = \overline{1,k})$, которые необходимо разделить на множества $V_j = (x^i, i \in I_j)$, $j = \overline{1,M}$, соответствующих одномодальным фрагментам плотности вероятности $p(x)$. Количество $M$ компактных групп наблюдений неизвестно.

Пусть в результате использования предложенной выше методики декомпозиции исходной информации получены данные $\bar{V} = (z_1^i, \ldots, z_k^i, \bar{P}^i, i = \overline{1,N})$, составленные из значений центров $z_v^i$, $v = \overline{1,k}$ интервалов (элементов) дискретизации $S^i$ пространства признаков $x$ анализируемых объектов и частот $\bar{P}^i$ принадлежности наблюдений выборки $V$ элементам $S^i$, $i = \overline{1,n}$. Тем самым для преодоления проблемы больших объёмов статистических данных плотность вероятности $p(x)$ заменяется на её оценку типа гистограммы.

Предлагаемый алгоритм автоматической классификации основан на выполнении следующих действий [6]:

1. Провести анализ массива данных $\bar{V}$ и исключить информацию элементов $S^i$, для которых $\bar{P}^i = 0$. Полученный массив преобразованных данных обозначим через $\tilde{V}$, а количество их элементов – как $\tilde{N}$. Множество их номеров обозначим через $\tilde{I}$.

2. Обнаружить элемент $S^q$ из $\tilde{V}$ с максимальной частотой

$$\bar{P}^q = \max_{i=\overline{1,\tilde{N}}} \bar{P}^i,$$

который отнести к классу $\Omega_1$.

3. Множество смежных с $S^q$ элементов $S(q) = (S^i, i \in I_1(q))$ будут отнесены к первому классу, так как априори значения $\bar{P}^q > \bar{P}^i$, $i \in I_1(q) = I_1^1$. Под смежными к $S^q$ понимаются элементы $S^i$, координаты которых удовлетворяют условиям:

$$\left| z_v^i - z_v^q \right| = \beta_v, \quad v = \overline{1,k}, \quad i \in \tilde{I}, \quad i \neq q,$$

где $\beta_v$ – длина интервала дискретизации по признаку $x_v$, $v = \overline{1,k}$.

Элементы, принадлежащие множеству $S(q)$, относятся к классу $\Omega_1$ и исключаются из множества $S^i$, $i \in \tilde{I}$, при последующей их идентификации.

4. Каждый элемент из множества $S^i \in S(q)$ является одним из центров для идентификации к классу $\Omega_1$ остальных ситуаций $\tilde{I} \setminus I_1(q)$.

По аналогии с пунктом 3 из множества $S(q)$ выбрать элемент $S^t$ и провести идентификацию к классу $\Omega_1$ смежных с ним элементов по правилу: элемент $S^i$ относится к классу $\Omega_1$, если

$$\left|z_v^i - z_v^t\right| = \beta_v,\ v = \overline{1,k},\ \text{а}\ \overline{P}^t > \overline{P}^i,\ i \in \left(\tilde{I} \setminus I_1(q) = I_1^1\right).$$

При соблюдении этих условий элемент $S^i$ относится к множеству $S_1^2$ в качестве центра для последующей идентификации.

5. Повторить этап 4 для всех элементов $S^i$, $i \in \left(\tilde{I} \setminus I_1(q) = I_1^1\right)$. В результате получим множество элементов $S_1^2$, отнесённых к первому классу $\Omega_1$ на данном этапе их идентификации. Множество номеров элементов $S_1^2$ обозначим через $I_1^2$.

6. Следуя предложенной методике, на $r$-м этапе обнаружения элементов, принадлежащих первому классу, осуществить их идентификацию по правилу:

$$\begin{aligned} &S^i \in \Omega_1, \text{ если} \left|z_v^i - z_v^j\right| = \beta_v, v = \overline{1,k}, \\ &\text{а}\ \overline{P}^j > \overline{P}^i, i \in \tilde{I} \setminus \bigcup_{t=1}^{r-1} I_1^t, j \in I_1^{r-1}. \end{aligned} \quad (6)$$

Процесс автоматической классификации в соответствии с этапом 6 продолжается до тех пор, пока при некотором значении $r$ условие (6) не будет выполняться. Первый класс образуют множество элементов

$$S^i,\ i \in \bigcup_{t=1}^{r-1} I_1^t.$$

7. Обнаружение множества элементов дискретизации, принадлежащих второму классу. Для этого по аналогии с этапами 2 – 6 провести анализ множества элементов

$$S^i,\ i \in \tilde{I} \setminus \bigcup_{t=1}^{r-1} I_1^t.$$

8. Процесс автоматической классификации продолжается до полного разбиения множества элементов

$$S^i,\ i \in \tilde{I} \setminus \bigcup_{t=1}^{r-1} I_1^t$$

на компактные группы элементов в соответствии с принятым определением класса.

Рассмотрим предложенный подход к решению задачи автоматической классификации для одномерного случая при $k = 1$, результаты дискретизации для которого представлены на рис. 1.

В результате дискретизации области значений случайной величины $x$ исходная выборка $V = (x^i, i = \overline{1,n})$ преобразуется в массив данных $\tilde{V} = (z^j, \overline{P}^j, i \in \tilde{I})$, для которых $\overline{P}^j \neq 0$. В соответствии с этапом 2 непараметрического алгоритма автоматической классификации к классу $\Omega_1$ будет отнесён интервал дискретизации $S^9$ с параметрами $(z^9, \overline{P}^9)$, который соответствует моде $\overline{p}_1(x)$. Тогда на этапе 3 классификации к первому классу $\Omega_1$ однозначно будут отнесены интервалы дискретизации $S^8$ и $S^{10}$ с параметрами $(z^8, \overline{P}^8)$, $(z^{10}, \overline{P}^{10})$ соответственно. Данное утверждение основывается на справедливости неравенств $\overline{P}^9 > \overline{P}^8$ и $\overline{P}^9 > \overline{P}^{10}$. Эти интервалы $S^8$ и $S^{10}$ образуют множество $S(9)$. Следуя этапу 4 алгоритма классификации, для анализа выбирается, например, интервал $S^8 \in S(9)$. В соответствии с решением алгоритма классификации этого этапа интервал $S^7$ будет отнесён к классу $\Omega_1$, так как справедливо соотношение $\overline{P}^8 > \overline{P}^7$. По аналогии относительно $S^{10} \in S(9)$ интервал $S^{11}$ будет отнесён также к классу $\Omega_1$, так как $\overline{P}^{10} > \overline{P}^{11}$. Далее проводится анализ интервалов $S(7)$ и $S(11)$, которые содержат только по одному интервалу $S^6$ и $S^{12}$ соответственно. Нетрудно заметить, что интервал $S^6$ не будет отнесён к первому классу $\Omega_1$, так как выполняется неравенство $\overline{P}^6 > \overline{P}^7$. Интервал $S^{12}$ будет отнесён с первому классу, потому что справедливо соотношение $\overline{P}^{11} > \overline{P}^{12}$. В данном примере к первому классу будут отнесены интервалы $S^j$, $j = \overline{7,12}$.

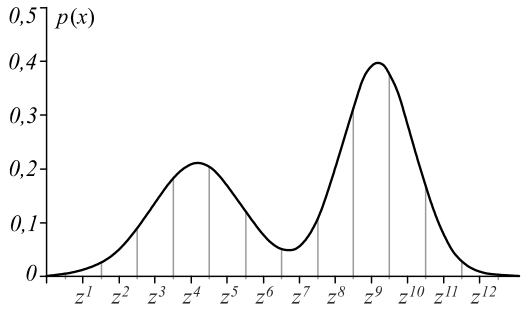


*Рис. 1. Графическая иллюстрация результатов дискретизации V области значений случайной величины x ($z^j$, j=1, 12 – центры интервалов дискретизации)*

Для обнаружения класса $\Omega_2$ необходимо из оставшегося массива данных $V$ выбрать интервал $S^4$ с максимальной частотой встречаемости $\overline{P}^4$ случайной величины из исходной выборки, и описанный выше процесс классификации повторяется. В результате обнаруживаются интервалы дискретизации $S^j$, $j = \overline{1,6}$, принадлежащие классу $\Omega_2$.

Нетрудно заметить, что основу предлагаемой процедуры классификации составляют оценка близости центров многомерных интервалов дискретизации и соотношений между их частотами. При этом осуществляется выделение классов, соответствующих одномодальным фрагментам совместной плотности вероятности анализируемых случайных величин.

Вычислительную эффективность предложенного алгоритма автоматической классификации в условиях больших объёмов статистических данных дополнительно можно повысить за счёт организации многопоточных вычислений при программной реализации процесса классификации. Для работы многопоточной обработки данных была использована встроенная в

язык C++ библиотека «Thread», из которой был использован класс «Std::thread». Сравнительные тесты для проверки эффективности многопоточных вычислений выполнялись на компьютере с процессором Intel® Core™ i5-6200U CPU@2,4 GHz (2 ядра, 4 потока).

По результатам вычислительных экспериментов многопоточный вариант программы автоматической классификации имеет двухкратное преимущество по времени по сравнению с однопоточным вариантом.

### *Оценивание состояний темнохвойных древостоев, повреждённых сибирским шелкопрядом, по данным дистанционного зондирования*

На исследуемой территории Ирбейского района Красноярского края преобладают пихтовые и кедровые древостои на высотах 300 – 1600 метров над уровнем моря. В период массового размножения сибирского шелкопряда на этой территории в 2018 – 2019 гг. погибло более 32 тысяч гектаров древостоев [12, 13].

Исходная информация сформирована 9 сентября 2019 года по данным дистанционного зондирования с помощью аппарата Landsat-8. Снимок получен с геопортала Earth Explorer, из которого вырезан тестовый участок в 11 тысяч гектар (рис. 2*а*). Он определялся 123134 пикселями. Каждый пиксель характеризовался семью спектральными признаками $x=(x_1,\ldots,x_7)$, которым соответствуют следующие длины волн (нанометры): 433 – 453 ($x_1$), 450 – 515 ($x_2$), 525 – 600 ($x_3$), 630 – 680 ($x_4$), 845 – 885 ($x_5$), 1560 – 1660 ($x_6$), 2100 – 2300 ($x_7$). Полученные данные подвергались атмосферной коррекции с помощью алгоритмов Land Surface Reflectance Code.

Для обнаружения компактных групп наблюдений в пространстве спектральных признаков $x=(x_1,\ldots,x_7)$ использовался предлагаемый непараметрический алгоритм автоматической классификации. Его программная реализация Autoclass 2.0 позволяет загружать изображения в формате GeoTIFF, производить классификацию с заданными параметрами и представлять сформированное классификационное изображение для дальнейшей геообрабоки. Рассматривались два варианта анализа исходного снимка без его топографической нормализации (рис. 2*а*) и с топографической нормализацией (рис. 3*а*). Топографическая нормализация снимка методом C-коррекции и использованием цифровой модели рельефа SRTM 1 arc с пространственным разрешением 30 метров значительно уменьшает эффект разной освещённости и выравнивает яркости однотипных поверхностей, что влияет на результаты автоматической классификации.

Применение программы Autoclass 2.0 при топографической нормализации (рис. 3*б*) сопровождается сокращением количества классов с $M=18$ до 11 по сравнению с условиями рис. 2*б*. Обнаруженные классы соответствуют лесным массивам с различной степенью поражения сибирским шелкопрядом, усохшим и лиственным древостоям, вырубкам, травяно-кустарниковым сообществам и заболоченным участкам.

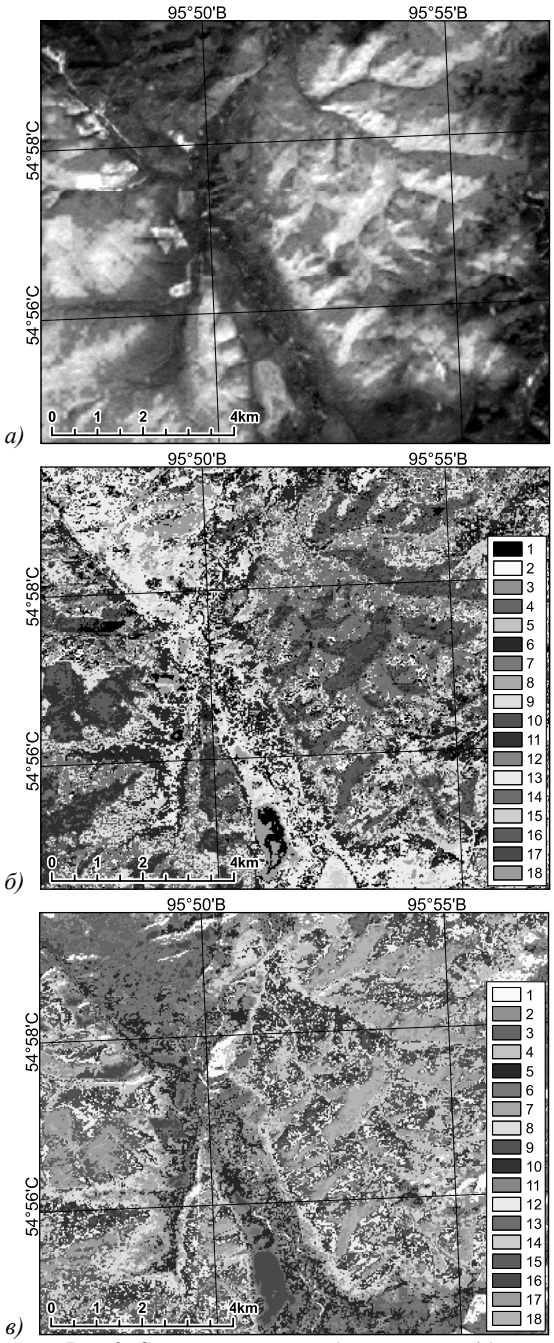


*Рис. 2. Сопоставление исходного снимка (а) без топографической нормализации и результаты автоматической классификации непараметрическим алгоритмом (б) и методом ISODATA (в)*

Результаты автоматической классификации в указанных условиях методом ISODATA средствами программного пакета Erdas Imagine приведены на рис. 2*в* и рис. 3*в*. При этом количество классов устанавливалось равным количеству классов, обнаруженных непараметрическим алгоритмом автоматической классификации.

Большую часть исследуемой территории представляют лесные массивы усохших темнохвойных древостоев, повреждённых сибирским шелкопрядом. Участки, близкие к правильной форме, соответствуют вырубкам. Светлыми тонами на рисунках показаны участки лиственных древостоев. Фоном, близким к

красному, определены участки темнохвойных древостоев, повреждённых сибирским шелкопрядом. Экспертный анализ показал, что результаты классификации указанными методами сопоставимы. Они хорошо выделяют зоны повреждённых древостоев на освещённых склонах и менее успешно – на затемнённых участках, что указывает на целесообразность предварительной топографической нормализации снимков до процедуры автоматической классификации. Оба метода классификации не позволяют отделить свежие вырубки от других поверхностей, таких как травянисто-кустарниковые сообщества и заболоченности. Однако применение непараметрического алгоритма позволяет получить более чёткие контуры вырубок (рис. 3).

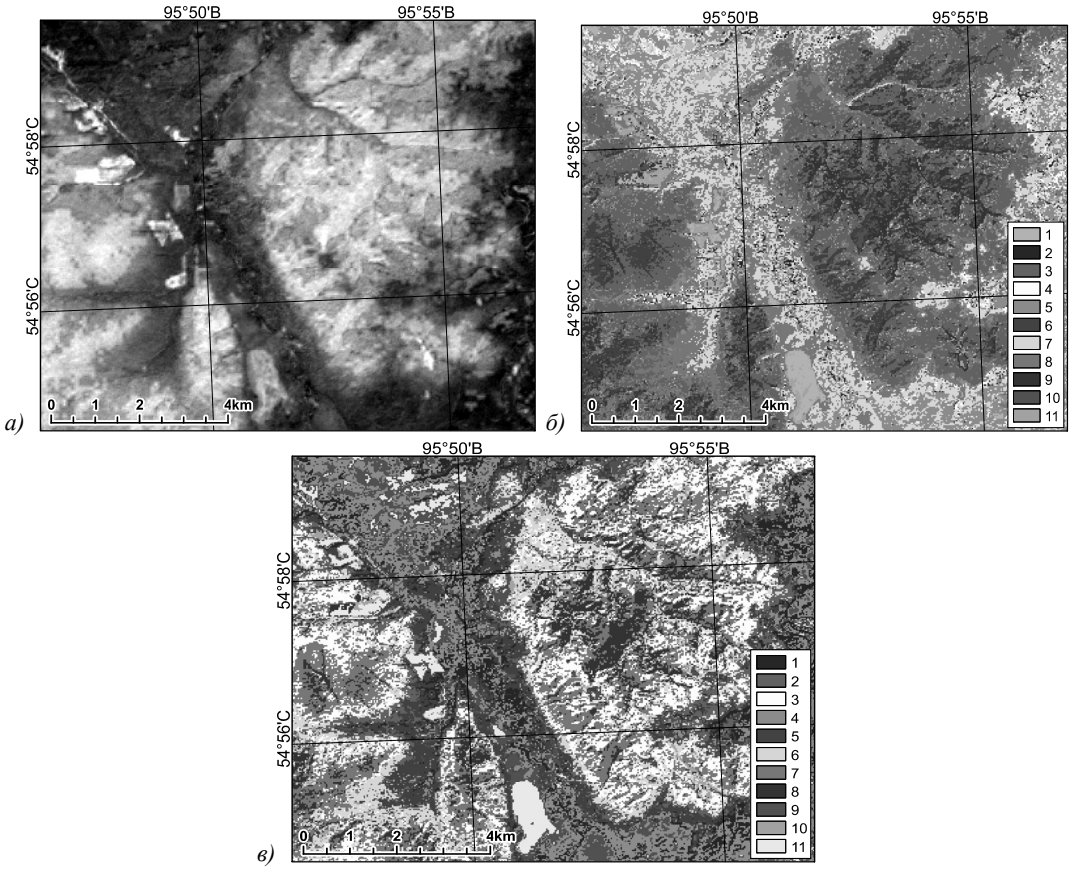


*Рис. 3. Сопоставление исходного снимка (а) с топографической нормализацией и результаты автоматической классификации непараметрическим алгоритмом (б) и методом ISODATA (в)*

### *Заключение*

Непараметрический алгоритм автоматической классификации статистических данных большого объёма основан на их «сжатии» путём декомпозиции многомерного пространства признаков исследуемых объектов. Полученная информация позволяет осуществить синтез регрессионной оценки плотности вероятности, асимптотические свойства которой определяют количество интервалов дискретизации области значений случайных величин. На этой основе формируется процедура автоматической классификации статистических данных, которая учитывает близость центров многомерных интервалов дискретизации и соотношения между частотами попадания случайных величин в эти интервалы. Вычислительная эффективность непараметрического алгоритма автоматической классификации повышается в два раза при использовании многопоточной технологии обработки данных при его программной реализации. Результаты исследования подтверждаются применением непараметрического алгоритма автоматической классификации при обработке спектральных данных дистанционного зондирования лесных массивов, повреждённых сибирским шелкопрядом. Топографическая нормализация исходного снимка позволяет повысить эффективность обнаружения состояний повреждённых лесных древостоев. Результаты автоматической классификации непараметрическим алгоритмом и методом программного продукта Erdas Imagine сопоставимы.

### *Благодарности*

Исследование выполнено при финансовой поддержке РФФИ, Правительства Красноярского края и Красноярского краевого фонда науки в рамках научного проекта № 20-41-240001.

### *Литература*

1. **Дорофеюк, А.А.** Алгоритмы автоматической классификации (обзор) / А.А. Дорофеюк // Автоматика и телемеханика. – 1971. – № 12. – С. 78-113.
2. **Дорофеюк, А.А.** Методология экспертно-классификационного анализа в задачах управления и обработки

сложноорганизованных данных (история и перспективы развития) / А.А. Дорофеюк // Проблемы управления. – 2009. – № 3(1). – С. 19-28.
3. **Цыпкин, Я.З.** Основы теории обучающихся систем / Я.З. Цыпкин. – М.: Наука, 1970. – 252 с.
4. **Васильев, В.И.** Особенности алгоритмов самообучения и кластеризации / В.И. Васильев, С.Н. Эш // Управляющие системы и машины. – 2011. – № 3. – С. 3-9.
5. **Лапко, А.В.** Непараметрический алгоритм автоматической классификации в условиях статистических данных большого объема / А.В. Лапко, В.А. Лапко // Информатика и системы управления. – 2018. – Т. 57, № 3. – С. 59-70. – DOI: 10.22250/isu.2018.57.59-70.
6. **Лапко, А.В.** Непараметрический алгоритм выделения классов, соответствующих одномодальным фрагментам плотности вероятности многомерных случайных величин / А.В. Лапко, В.А. Лапко, С.Т. Им, В.П. Тубольцев, В.Л. Авдеенок // Автометрия. – 2019. – Т. 55, № 3. – С. 22-30. – DOI: 10.15372/AUT20190303.
7. **Лапко, А.В.** Регрессионная оценка многомерной плотности вероятности и её свойства / А.В. Лапко, В.А. Лапко // Автометрия. – 2014. – Т. 50, № 2. – С. 50-56.
8. **Лапко, А.В.** Оптимальный выбор количества интервалов дискретизации области изменения одномерной случайной величины при оценивании плотности вероятности / А.В. Лапко, В.А. Лапко // Измерительная техника. – 2013. – № 7. – С. 24-27.
9. **Лапко, А.В.** Выбор оптимального количества интервалов дискретизации области значений двухмерной случайной величины / А.В. Лапко, В.А. Лапко // Измерительная техника. – 2016. – № 2. – С. 14-17.
10. **Лапко, А.В.** Метод дискретизации области значений многомерной случайной величины / А.В. Лапко, В.А. Лапко // Измерительная техника. – 2019. – № 1. – С. 16-20. – DOI: 10.32446/0368-1025it.2019-1-16-20.
11. **Лапко, А.В.** Оценивание интеграла от квадрата плотности вероятности одномерной случайной величины / А.В. Лапко, В.А. Лапко // Измерительная техника. – 2020. – № 7. – С. 22-28. – DOI: 10.32446/0368-1025it.2020-7-22-28.
12. **Kharuk, V.I.** Climate-induced northerly expansion of Siberian silkmoth range / V.I. Kharuk, S.T. Im, K.J. Ranson, M.N. Yagunov // Forests. – 2017. – Vol. 8, Issue 8. – 301. – DOI: 10.3390/f8080301.
13. **Kharuk, V.I.** Siberian silkmoth outbreaks surpassed geoclimatic barrier in Siberian Mountains / V.I. Kharuk, S.T. Im, V.V. Soldatov // Journal of Mountain Science. – 2020. – Vol. 17. – P. 1891-1900. – DOI: 10.1007/s11629-020-5989-3.

### *Сведения об авторах*

**Зеньков Игорь Владимирович**, 1963 года рождения, в 1985 г. окончил Красноярский институт цветных металлов по специальности «Технология и комплексная механизация открытой разработки месторождений полезных ископаемых», доктор технических наук, профессор, профессор кафедры систем автоматики, автоматизированного управления и проектирования Сибирского федерального университета, ведущий научный сотрудник Красноярского филиала Федерального исследовательского центра информационных и вычислительных технологий. Область научных интересов: решение задач горнодобывающей промышленности с использованием ресурсов дистанционного зондирования; информационное обеспечение мониторинга технологических, логистических параметров предприятий горной промышленности; дистанционное зондирование.
E-mail: *zenkoviv@mail.ru* .

**Лапко Александр Васильевич**, 1949 года рождения, в 1971 году окончил Фрунзенский политехнический институт по специальности «Автоматика и телемеханика», доктор технических наук, профессор, заслуженный деятель науки РФ, главный научный сотрудник Института вычислительного моделирования Сибирского отделения Российской академии наук, профессор кафедры космических средств и технологий Сибирского государственного университета науки и технологий имени академика М.Ф. Решетнева. Область научных интересов: непараметрическая статистика; распознавание образов и анализ изображений; моделирование и оптимизация неопределённых систем, дистанционное зондирование. E-mail: *lapko@icm.krasn.ru* .

**Лапко Василий Александрович**, 1974 года рождения, в 1996 году окончил Красноярский государственный технический университет по специальности «Управление и информатика в технических системах», доктор технических наук, профессор, ведущий научный сотрудник Института вычислительного моделирования Сибирского отделения Российской академии наук, заведующий кафедрой космических средств и технологий Сибирского государственного университета науки и технологий имени академика М.Ф. Решетнева. Область научных интересов: непараметрическая статистика; распознавание образов и анализ изображений; моделирование неопределённых систем, дистанционное зондирование. E-mail: *valapko@yandex.ru* .

**Им Сергей Тхекдеевич**, 1979 года рождения, в 2001 году окончил Красноярский государственный технический университет по специальности «Информационные системы в геоинформационных системах», кандидат технических наук, доцент кафедры географии Сибирского федерального университета, ведущий научный сотрудник Института леса имени В.Н. Сукачева Сибирского отделения Российской академии наук, доцент кафедры космических средств и технологий Сибирского государственного университета науки и технологий имени академика М.Ф. Решетнева. Основная область научных интересов: исследование пространственно-временной

динамики и состояния лесных территорий на основе данных дистанционного зондирование Земли и геоинформационных систем. E-mail: stim@ksc.krasn.ru .

**Тубольцев Виталий Павлович**, 1998 год рождения, в 2020 году окончил Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева по направлению 21.03.03 «Геодезия и дистанционное зондирование» со степенью бакалавра. Поступил на первый курс магистратуры Сибирского государственного университета науки и технологий имени академика М.Ф. Решетнева по направлению 09.04.02 «Информационные системы и технологии» по профилю «Информационные системы обработки данных дистанционного зондирования». Область научных интересов: разработка информационных средств, непараметрические системы классификации, быстрые алгоритмы оптимизации решающих правил, обработка данных дистанционного зондирования. E-mail: vitalya.98@mail.ru .

**Авдеенок Валерий Леонидович**, 1998 год рождения, в 2020 году окончил Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева по направлению 21.03.03 «Геодезия и дистанционное зондирование» со степенью бакалавра. Поступил на первый курс магистратуры Сибирского государственного университета науки и технологий имени академика М.Ф. Решетнева по направлению 09.04.02 «Информационные системы и технологии» по профилю «Информационные системы обработки данных дистанционного зондирования». Область научных интересов: разработка информационных средств, параллельные вычислительные технологии, непараметрические системы классификации, дистанционное зондирование.
E-mail: avdeyonok@gmail.com .

*ГРНТИ: 28.23.15*
*Поступила в редакцию 21 августа 2020 г. Окончательный вариант – 3 декабря 2020 г.*

# A nonparametric algorithm for automatic classification of large multivariate statistical data sets and its application

*I.V. Zenkov [1,5], A.V. Lapko [2,4], V.A. Lapko [2,4], S.T. Im [1,3,4], V.P. Tuboltsev [4], V.L. Avdeenok [4]*
*[1]Siberian Federal University,*
*660041, Krasnoyarsk, Russia, Svobodny Av. 79,*
*[2]Institute of Computational Modelling SB RAS,*
*660036, Krasnoyarsk, Russia, Akademgorodok 50,*
*[3]Sukachev Institute of Forest SB RAS,*
*660036, Krasnoyarsk, Russia, Akademgorodok 50,*
*[4]Reshetnev Siberian State University of Science and Technology,*
*660037, Krasnoyarsk, Russia, Krasnoyarsky Rabochy Av. 31,*
*[5]Krasnoyarsk Branch of the Federal Research Center for Information and Computational Technologies,*
*660049, Krasnoyarsk, Russia, Mira Av. 53*

***Abstract***

A nonparametric algorithm for automatic classification of large statistical data sets is proposed. The algorithm is based on a procedure for optimal discretization of the range of values of a random variable. A class is a compact group of observations of a random variable corresponding to a unimodal fragment of the probability density. The considered algorithm of automatic classification is based on the «compression» of the initial information based on the decomposition of a multidimensional space of attributes. As a result, a large statistical sample is transformed into a data array composed of the centers of multidimensional sampling intervals and the corresponding frequencies of random variables. To substantiate the optimal discretization procedure, we use the results of a study of the asymptotic properties of a kernel-type regression estimate of the probability density. An optimal number of sampling intervals for the range of values of one- and two-dimensional random variables is determined from the condition of the minimum root-mean square deviation of the regression probability density estimate. The results obtained are generalized to the discretization of the range of values of a multidimensional random variable. The optimal discretization formula contains a component that is characterized by a nonlinear functional of the probability density. An analytical dependence of the detected component on the antikurtosis coefficient of a one-dimensional random variable is established. For independent components of a multidimensional random variable, a methodology is developed for calculating estimates of the optimal number of sampling intervals for random variables and their lengths. On this basis, a nonparametric algorithm for the automatic classification is developed. It is based on a sequential procedure for checking the proximity of the centers of multidimensional sampling intervals and relationships between frequencies of the membership of the random variables from the original sample of these intervals. To further increase the computational efficiency of the proposed automatic classification algorithm, a multithreaded method of its software implementation is used. The practical significance of the developed algorithms is confirmed by the results of their application in processing remote sensing data.

*Keywords:* automatic classification algorithm, multidimensional histogram, regression probability density estimate, discretization of the range of values of a random variable, large samples, antikurtosis coefficient, remote sensing data.

*Citation:* Zenkov IV, Lapko AV, Lapko VA, Im ST, Tuboltsev VP, Avdeenok VL. A nonparametric algorithm for automatic classification of large multivariate statistical data sets and its application. Computer Optics 2021; 45(2): 253-260. DOI: 10.18287/2412-6179-CO-801.

*Acknowledgements*: The research was funded by RFBR, Krasnoyarsk Territory and Krasnoyarsk Regional Fund of Science, project number 20-41-240001.

***References***

[1] Dorofeyuk AA. Algorithms of automatic classification (review) [In Russian]. Automation and Remote Control 1971; 12: 78-113.

[2] Dorofeyuk AA. Methodology of expert classification analysis in the management and processing of complex data (history and prospects of development) [In Russian]. Control Sciences 2009; 3(1): 19-28.

[3] Tsypkin YaZ. Fundamentals of the theory of learning systems [In Russian]. Moscow: "Nauka" Publisher; 1970.

[4] Vasil'ev VI, Ehsh SN. Features of self-learning algorithms and clustering [In Russian]. Control Systems and Machines 2011; 3: 3-9.

[5] Lapko AV, Lapko VA. Nonparametric algorithm of automatic classification under conditions of large-scale statistical data [In Russian]. Informatika i Sistemy Upravleniya 2018; 57(3): 59-70. DOI: 10.22250/isu.2018.57.59-70.
[6] Lapko AV, Lapko VA, Im ST, Tuboltsev VP, Avdeenok VL. Nonparametric algorithm of identification of classes corresponding to single-mode fragments of the probability density of multidimensional random variables. Optoelectronics, Instrumentation and Data Processing 2019; 55(3): 230-236. DOI: 10.3103/S8756699019030038.
[7] Lapko AV, Lapko VA. Regression estimate of the multidimensional probability density and its properties. Optoelectronics, Instrumentation and Data Processing 2014; 50(2): 148-153. DOI: 10.3103/S875669901402006X.
[8] Lapko AV, Lapko VA. Optimal selection of the number of sampling intervals in domain of variation of a one-dimensional random variable in estimation of the probability density. Measurement Techniques 2013; 56(7): 763-767. DOI: 10.1007/s11018-013-0279-x.
[9] Lapko AV, Lapko VA. Selection of the optimal number of intervals sampling the region of values of a two-dimensional random variable. Measurement Techniques 2016; 59(2): 122-126. DOI: 10.1007/s11018-016-0928-y.
[10] Lapko AV, Lapko VA. Discretization method for the range of values of a multi-dimensional random variable. Measurement Techniques 2019; 62(1): 16-22. DOI: 10.1007/s11018-019-01579-0.
[11] Lapko AV, Lapko VA. Estimating the integral of the square of the probability density of a one-dimensional random variable. Measurement Techniques 2020; 63: 534-542. DOI: 10.1007/s11018-020-01820-1.
[12] Kharuk VI, Im ST, Ranson KJ, Yagunov MN. Climate-induced northerly expansion of Siberian silkmoth range. Forests 2017; 8(8): 301. DOI: 10.3390/f8080301.
[13] Kharuk VI, Im ST, Soldatov VV. Siberian silkmoth outbreaks surpassed geoclimatic barrier in Siberian Mountains. Journal of Mountain Science 2020; 17: 1891-1900. DOI: 10.1007/s11629-020-5989-3.

### *Authors' information*

**Igor Vladimirovich Zenkov** (b. 1963), graduated from Krasnoyarsk Institute of Non-ferrous Metals on speciality "Technology and Complex Mechanization of Opencast Mining of Mineral Deposits" in 1985. Doctor of Science in Technology, professor, professor of the Automation Systems, Automated Control and Design department at Siberian Federal University; leading researcher at the Krasnoyarsk Branch of the Federal Research Center for Information and Computational Technologies. Research interests: solving problems in the mining industry using remote sensing resources; information support for monitoring technological, logistic parameters of mining enterprises; remote sensing. E-mail: *zenkoviv@mail.ru* .

**Alexander Vasilievich Lapko** (b. 1949), graduated from Frunze Polytechnic Institute on speciality "Automation and Telemechanics" in 1971. Doctor of Science in Technology, professor, honored worker of science of the Russian Federation, chief researcher of the Institute of Computational Modeling of the Siberian Branch of the Russian Academy of Sciences; Professor of Space Facilities and Technologies department of the Reshetnev Siberian State University of Science and Technology. Research interests: nonparametric statistics; pattern recognition and image analysis; modeling and optimization of uncertain systems; remote sensing. E-mail: *lapko@icm.krasn.ru* .

**Vasiliy Aleksandrovich Lapko** (b. 1974), graduated from Krasnoyarsk State Technical University on speciality "Management and Informatics in Technical Systems" in 1996. Doctor of Science in Technology, professor, leading researcher at the Institute of Computational Modeling of the Siberian Branch of the Russian Academy of Sciences; Head of Space Facilities and Technologies department of the Reshetnev Siberian State University of Science and Technology. Research interests: nonparametric statistics; pattern recognition and image analysis; modeling of uncertain systems; remote sensing. E-mail: *valapko@yandex.ru* .

**Sergei Thekdeyevich Im** (b. 1979), graduated from Krasnoyarsk State Technical University on speciality "Information Systems in Geoinfromation Systems" in 2001. Candidate of Sciences in Technology, Docent, Leading Researcher at the Institute of Forest of the Siberian Branch of the Russian Academy of Sciences; Docent of the Space Facilities and Technologies department of the Reshetnev Siberian State University of Science and Technology. Research interests: analysis of spatial-temporal dynamics and monitoring of forest based on the remote sensing data and geoinformation systems. E-mail: *stim@ksc.krasn.ru* .

**Vitaly Pavlovich Tuboltsev** (b. 1998), graduated from the Reshetnev Siberian State University of Science and Technology on speciality 21.03.03 «Geodesy and Remote Sensing» in 2020. First-year graduate student at the Reshetnev Siberian State University of Science and Technology on speciality 09.04.02 «Information Systems and Technologies». Research interests: development of information tools, nonparametric classification systems, fast optimization algorithms for decision rules, remote sensing processing. E-mail: *vitalya.98@mail.ru* .

**Valery Leonidovich Avdeenok** (b. 1998), graduated from the Reshetnev Siberian State University of Science and Technology on speciality 21.03.03 «Geodesy and Remote Sensing» in 2020. First-year graduate student at the Reshetnev Siberian State University of Science and Technology on speciality 09.04.02 «Information Systems and Technolo-

gies». Research interests: development of information tools, parallel computing technologies, nonparametric classification systems, remote sensing. E-mail: *avdeyonok@gmail.com* .

*Received August 21, 2020. The final version – December 3, 2021.*