УДК: 519.7 + 004.93

© 2019 г. **А.В. Лапко** [1,3], д-р техн. наук,
В.А. Лапко [1,3], д-р техн. наук,
С.Т. Им [2,3], канд. техн. наук,
В.П. Тубольцев [3],
В.Л. Авдеенок [3],
А.В. Бахтина [3]
([1]Институт вычислительного моделирования СО РАН, Красноярск,
[2]Институт леса им. В.Н. Сукачева СО РАН, Красноярск,
[3]Сибирский государственный университет науки и технологий
имени академика М.Ф. Решетнева, Красноярск)

ПРОГРАММНЫЕ СРЕДСТВА РЕАЛИЗАЦИИ НЕПАРАМЕТРИЧЕСКОГО АЛГОРИТМА АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ СТАТИСТИЧЕСКИХ ДАННЫХ БОЛЬШОГО ОБЪЕМА*

Рассматриваются программные средства, реализующие непараметрический алгоритм автоматической классификации статистических данных большого объема. Класс характеризуется одномодальным фрагментом плотности вероятности. Предлагаемый алгоритм автоматической классификации основан на декомпозиции исходных данных. Результаты декомпозиции образуют множество центров многомерных интервалов и соответствующие им частоты встречаемости значений случайных величин. На основе полученной информации обнаруживаются классы. Полученные непараметрические алгоритмы имеют важное значение при обработке данных дистанционного зондирования.

Ключевые слова: программные средства, автоматическая классификация, многомерная гистограмма, распознавание образов, выборки большого объема, дискретизация области значений многомерных случайных величин, данные дистанционного зондирования.

DOI: 10.22250/isu.2019.61.81-87

* Исследование выполнено при финансовой поддержке РФФИ (грант № 18-01-00251).

Введение

Методы автоматической классификации являются основой решения задач обнаружения скрытых закономерностей, свойственных объектам различной природы, информация о которых доступна в массивах наблюдений их многомерных признаков. Систематизация алгоритмов автоматической классификации рассмотрена в работах [1, 2]. Активно развивающимся направлением разбиения статистических данных на множества компактных групп наблюдений являются непараметрические методы ядерного типа. Разработан ряд подобных алгоритмов автоматической классификации, обеспечивающих выделение множества наблюдений с одномодальными фрагментами плотностей вероятностей случайных величин [3, 4]. В данной работе рассматривается программная реализация указанных выше непараметрических алгоритмов автоматической классификации и результаты их применения при обработке данных дистанционного зондирования.

Методика синтеза непараметрического алгоритма автоматической классификации

Исходную выборку $V = \left(x^i, i = \overline{1, n}\right)$ большого объема n значений многомерной случайной величины $x = \left(x_v, v = \overline{1, k}\right)$ с плотностью вероятности $p(x_1, \ldots, x_k)$ преобразуем в массив данных $\overline{V} = \left(z_1^i, \ldots, z_k^i, \overline{P}^i, i = \overline{1, N}\right)$ по методике работы [4]. Элементы выборки $\overline{V}$ определяются значениями центров z_v^i, $v = \overline{1, k}$ интервалов дискретизации и соответствующих им частот $\overline{P}^i$, $i = \overline{1, n}$. Оптимальное количество интервалов дискретизации области значений многомерной случайной величины определяется формулой [5]

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

Коэффициент $\alpha(k) \le 1$ и его значения уменьшаются с ростом размерности k случайной величины. Значения $\alpha(k) = (2k-1)/k^2$.

Формула справедлива и для разных интервалов Δ_v значений случайных величин x_v, $v = \overline{1, k}$. Количество интервалов дискретизации N_v по каждому признаку x_v определяется соотношением $\min\limits_N \left| N_k^* - N^k \right|$ при $N_v = N$, $v = \overline{1, k}$, где N – целое число.

Рассмотрим предлагаемый подход к решению задачи автоматической классификации для одномерного случая при $k = 1$, результаты дискретизации

для которого представлены на рис. 1.

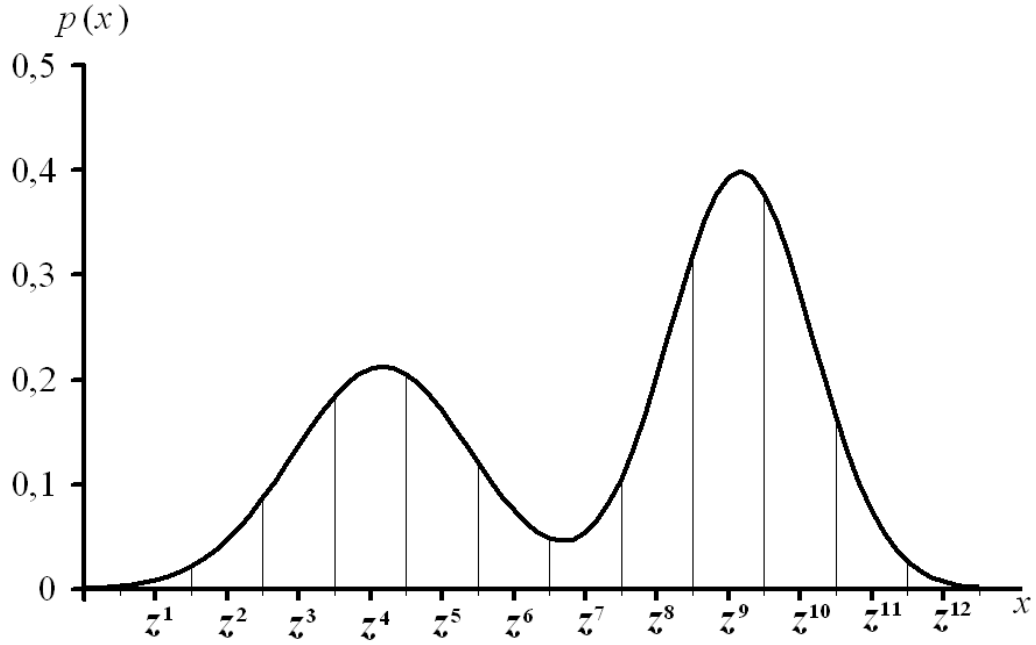

Рис. 1. Графическая иллюстрация результатов дискретизации $\widetilde{V}$ области значений случайной величины x (z^j, $j=\overline{1,12}$ – центры интервалов дискретизации).

В результате дискретизации области значений случайной величины x исходная выборка $V=\left(x^i, i=\overline{1,n}\right)$ преобразуется в массив данных $\widetilde{V}=\left(z^j, \overline{P}^j, i\in\widetilde{I}\right)$, для которых $\overline{P}^j\neq 0$. В соответствии с непараметрическим алгоритмом автоматической классификации к классу Ω_1 будет отнесено множество наблюдений из интервала дискретизации S^9 с параметрами $\left(z^9, \overline{P}^9\right)$. Тогда на следующем этапе классификации к первому классу Ω_1 однозначно будут отнесены множества наблюдений из интервалов дискретизации S^8 и S^{10} с параметрами $\left(z^8, \overline{P}^8\right)$, $\left(z^{10}, \overline{P}^{10}\right)$ соответственно. Это утверждение основывается на справедливости неравенств $\overline{P}^9>\overline{P}^8$ и $\overline{P}^9>\overline{P}^{10}$. Интервалы S^8 и S^{10} образуют множество $S(9)$. Согласно алгоритму классификации для последующего анализа выбирается, например, интервал $S^8\in S(9)$. В соответствии с решением этого этапа алгоритма классификации множество наблюдений из интервала S^7 будет отнесено к классу Ω_1, так как справедливо соотношение $\overline{P}^8>\overline{P}^7$. По аналогии относительно $S^{10}\in S(9)$ множество наблюдений из интервала S^{11} будут отнесены также к классу Ω_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 множество наблюдений из интервала S^6 не будет отнесено к первому классу Ω_1, так как

выполняется неравенство $\overline{P}^6 > \overline{P}^7$. Множество наблюдений из интервала S^{12} будет отнесено с первому классу, потому что справедливо соотношение $\overline{P}^{11} > \overline{P}^{12}$. В данном примере к первому классу будут отнесены множества наблюдений из интервалов S^j, $j = \overline{7,12}$.

Для обнаружения класса Ω_2 необходимо из оставшегося массива данных $\widetilde{V}$ 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 множества наблюдений из интервалов дискретизации S^j, $j = \overline{1,6}$, принадлежащих классу Ω_2.

Первые из обнаруженных классов будут содержать большее количество интервалов дискретизации, которые имеют тенденцию убывания по мере последовательного обнаружения классов. 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.

Программное обеспечение непараметрического алгоритма автоматической классификации

Для достижения цели работы разработаны программные средства, реализующие непараметрический алгоритм автоматической классификации в среде Visual Studio Community 2017. Программные средства реализуют следующие функциональные возможности:

1) первичная обработка статистических данных большого объема, обеспечивающая оценивание основных количественных характеристик законов распределения случайных величин (математическое ожидание и его доверительное оценивание, среднеквадратическое отклонение);

2) обнаружение классов статистических данных, соответствующих одномодальным фрагментам плотности вероятности многомерных случайных величин на основе непараметрического алгоритма автоматической классификации;

3) пространственное отображение результатов автоматической классификации и анализ их количественных характеристик, соответствующих им законов распределения.

Программные средства охватывают основные этапы обработки статистических данных большого объема в соответствии со следующей ее структурой (рис. 2).

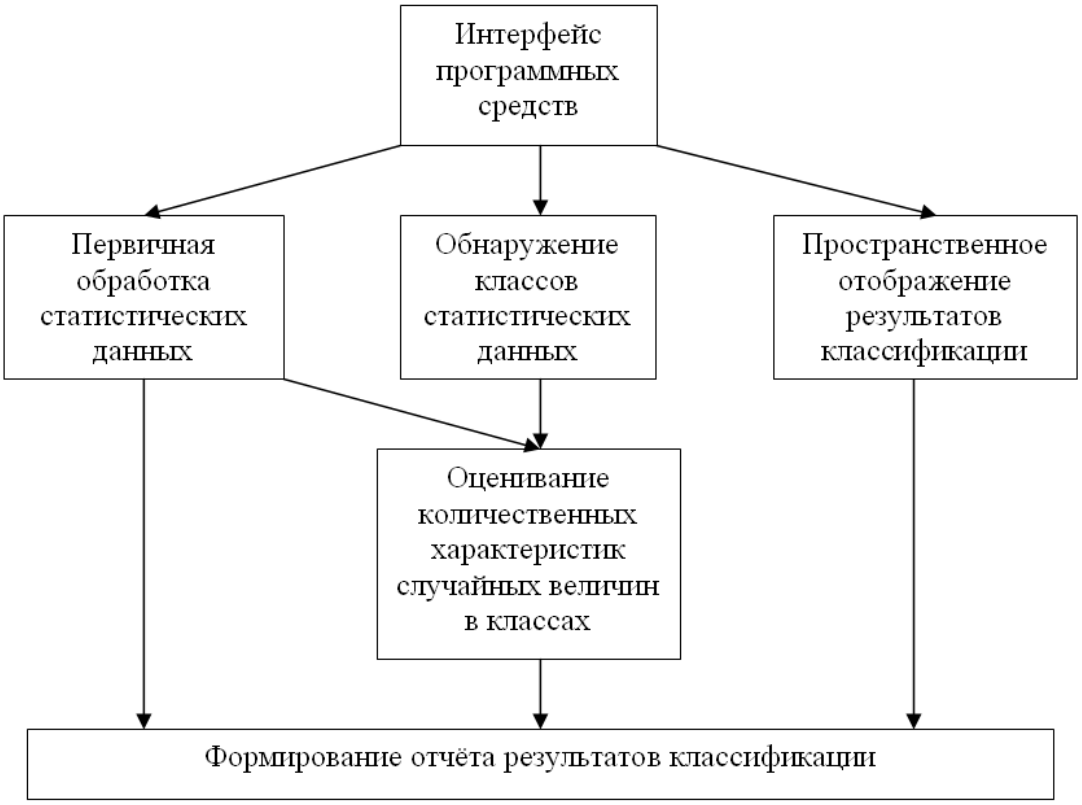

Рис. 2. Структура программных средств, реализующих непараметрический алгоритм автоматической классификации.

Интерфейс программы позволяет осуществлять эффективное управление исходными данными, последовательностью их обработки и представлением результатов решения функциональных задач.

Результаты автоматической классификации спектральных данных дистанционного зондирования

Исследуемая территория расположена в центре Восточно-Сибирского региона России, в центральной части Ангарского кряжа (рис. 3). Исходная информация формировалась по фрагменту спутниковой съемки Landsat 8 OLI с пространственным разрешением 30 метров. Размер фрагмента составляет 320×261 пиксель. Каждый пиксель характеризовался пятью спектральными признаками: синий x_1 (длина волны 0,433 – 0,453 мкм), зеленый x_2 (0,525 – 0,600 мкм), красный x_3 (0,630 – 0,680 мкм), ближний инфракрасный x_4 (0,845 – 0,885 мкм), ближний инфракрасный x_5 (1,560 – 1,660 мкм).

Для использования непараметрического алгоритма автоматической классификации определялось оптимальное количество интервалов дискрети-

зации для каждого спектрального канала 13, 13, 12, 10, 10 соответственно. При этом количество элементов дискретизации определялось значением 202800.

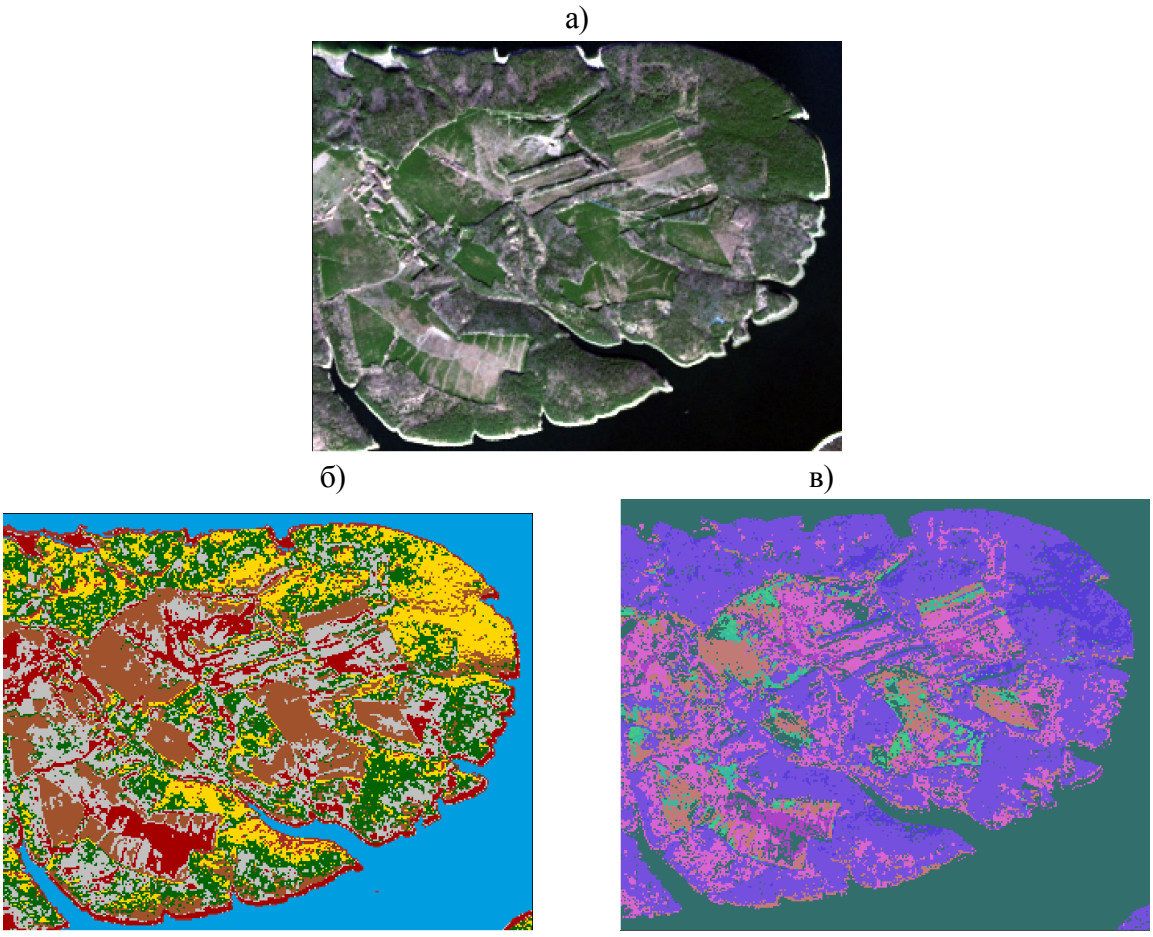

Рис. 3. Фрагмент спутниковой съемки Landsat 8 OLI (а).
Пространственное отображение результатов автоматической классификации,
полученных с использование программного продукта Erdas Imagine (б)
и непараметрического алгоритма (в) при количестве классов $M = 8$.

После удаления элементов с нулевыми частотами количество значимых элементов соответствовало значению 52, по которым формировались данные для автоматической классификации.

Заключение

Программные средства, реализующие непараметрический алгоритм автоматической классификации в условиях статистических данных большого объема, представлены в среде Visual Studio Community 2017. Их функциональные возможности обеспечивают обнаружение классов, соответствующих одномодальным фрагментам плотности вероятности случайных величин, оценивание их вероятностных характеристик и пространственное отображение результатов классификации.

Предлагаемый непараметрический алгоритм автоматической классификации, в условиях статистических данных большого объема, основан на их «сжатии» путем декомпозиции многомерного пространства признаков. В результате исходная выборка преобразуется в массив данных, составленный из центров многомерных интервалов дискретизации и соответствующих им частот принадлежности случайных величин. Основу процедуры классификации составляет проверка близости центров многомерных интервалов дискретизации и соотношений между их частотами. При этом выделяются классы, количество которых априори не задается. Результаты применения непараметрического алгоритма автоматической классификации и программного продукта Erdas Imagine при обработке данных дистанционного зондирования исследуемой территории визуально сопоставимы. Однако количество обнаруживаемых классов и распределения между ними спектральных данных при использовании сравниваемых методов могут быть разными. Это создает основу для планирования полевых исследований анализируемых территорий с целью обоснования результатов классификации.

ЛИТЕРАТУРА

1. *Дорофеюк А.А.* Алгоритмы автоматической классификации (обзор) // 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) // Проблемы управления. – 2009. – №3.1. – С. 19-28.
3. *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. DOI: 10.22250/isu.2018.57.59-70
4. *Лапко А.В., Лапко В.А., Им С.Т., Тубольцев В.П., В.А. Авдеенок* Непараметрический алгоритм выделения классов, соответствующих одномодальным фрагментам плотности вероятности многомерных случайных величин // Автометрия. – 2019. – Т.55. – №3. – С. 22-30. DOI: 10.15372/AUT20190303.
5. *Лапко А.В., Лапко В.А.* Метод дискретизации области значений многомерной случайной величины // Измерительная техника. – 2019. – №1. – С. 16-20. DOI: 10.32446/0368-1025it.2019-1-16-20.

E-mail:
Лапко Александр Васильевич – lapko@icm.krasn.ru;
Лапко Василий Александрович – lapko@icm.krasn.ru;
Им Сергей Тхекдеевич – stim@ksc.krasn.ru;
Тубольцев Виталий Павлович – vitalya.98@mail.ru;
Авдеенок Валерий Леонидович – avdeyonok@gmail.com;
Бахтина Анна Владлимировна – anna-denisyuk@yandex.ru.