

Непараметрический алгоритм автоматической классификации для тематической обработки данных дистанционного зондирования

И.А. Пестунов, Е.А. Будкина, Ю.Н. Синявский¹

¹*Институт вычислительных технологий СО РАН, Новосибирск, Россия
e-mail: pestunov@ict.nsc.ru*

The fast nonparametric algorithm of automatic classification for segmentation of the multi-spectral aerospace images is offered. The algorithm is based on formation of cell-like data structure in space of spectral features and realization on its basis of an iterative procedure of mean shift for searching local modes of a density function. The results of an experimental research of algorithm on real aerospace data are resulted.

Введение

В настоящее время при решении задач, связанных с исследованием природных ресурсов, тематическим картированием и оценкой экологического состояния территорий, широко используются аэрокосмические видеоданные, получаемые с помощью многозональных сканирующих систем. Для тематической классификации этих видеоданных в ряде специализированных пакетов (ERDAS Imagine, ENVI, IDRISI и др.) имеются соответствующие программные модули. Однако возможности получения качественных результатов тематической обработки во многом ограничиваются тем обстоятельством, что лежащие в основе этих программных средств алгоритмы классификации ориентированы либо на использование классифицированной обучающей выборки (КОВ), либо на необходимость задания пользователем ряда параметров, которые должны предопределять форму классов, их размер и количество. Но на практике априорная информация, необходимая для корректного задания значений параметров классов, как правило, отсутствует, а получение КОВ нередко сопряжено с неприемлемо большими экономическими затратами.

В работе предлагается непараметрический алгоритм автоматической классификации, для корректного применения которого не требуются ни КОВ, ни наличие каких-либо априорных сведений о характеристиках классов. Алгоритм опирается на клеточную структуру данных, формируемую в пространстве спектральных признаков, и реализованную на ее основе итеративную процедуру среднего сдвига, которая порождает естественное разбиение многозональных данных на классы [1,2].

Постановка задачи и метод ее решения

Предположим, что произведена k -зональная съемка участка местности, содержащего N элементов разрешения, тогда результат съемки можно представить в виде множества $X = \{x^{(i)} = (x_1^{(i)}, \dots, x_k^{(i)}) \in R^k, i = \overline{1, N}\}$, где $x_j^{(i)}$ – значение яркости i -го элемента

разрешения в j -ом диапазоне спектра ($j = \overline{1, k}$). Пусть каждый вектор $x^{(i)}$ – реализация k -мерного случайного вектора \mathbf{x} , плотность распределения которого $f(x)$, $x \in R^k$, неизвестна и нет какой-либо априорной информации о ее параметрическом виде. В этих условиях для оценивания плотности $f(x)$ в точке $x \in R^k$ целесообразно воспользоваться непараметрической парzenовской оценкой $\hat{f}_N(x)$, определяемой выражением

$$\hat{f}_N(x) = \frac{1}{Nh^k} \sum_{i=1}^N \Phi \left(\frac{x - x^{(i)}}{h} \right),$$

где $\Phi(x)$ – скалярная функция (ядро), удовлетворяющая следующим условиям:

$$1) \Phi(x) \geq 0 \quad \forall x \in R^k, \quad 2) \sup_{x \in R^k} \Phi(x) < \infty, \quad 3) \int_{R^k} \Phi(x) dx = 1, \quad 4) \lim_{\|x\| \rightarrow \infty} \|x\|^k \Phi(x) = 0.$$

Эта оценка является простой и, в отличие от гистограммной оценки и оценки k -ближайших соседей, обладает высокими асимптотическими свойствами. Она является несмещенной, состоятельной в среднеквадратическом смысле и равномерно сходящейся по вероятности при условии, что $\lim_{N \rightarrow \infty} h(N) = 0$, $\lim_{N \rightarrow \infty} Nh^k(N) < \infty$, $\lim_{N \rightarrow \infty} Nh^{2k}(N) = \infty$.

Пусть $m_h(x) = 1/n_x \sum_{x^{(i)} \in S_h(x)} x^{(i)}$ – выборочное среднее в точке $x \in R^k$. Здесь $S_h(x)$ – гипершар с центром в точке x и радиусом h , а n_x – количество точек множества X , содержащихся в $S_h(x)$. Тогда, согласно [1,2], разность $m_h(x) - x$ есть вектор среднего сдвига. Этот вектор интересен тем, что его направление совпадает с направлением градиента оценки $\hat{f}_N(x)$ в точке x , если в качестве ядра этой оценки использовать ядро Епанечникова, определяемое выражением

$$\Phi_E(x) = \begin{cases} \frac{1}{2} V_k^{-1} (k+2) (1 - x^T x), & \text{если } x^T x < 1 \\ 0, & \text{если } x^T x \geq 1, \end{cases}$$

где V_k – объем единичного k -мерного шара.

Повторяющиеся движения от точки $x \in R^k$ к ее выборочному среднему $m_h(x)$, затем от $x_1 = m_h(x)$ к $m_h(x_1)$ и т.д., пока на некотором шаге n значение $m_h(x_n)$ не будет равно $m_h(x_{n+1})$, называют алгоритмом среднего сдвига. Этот алгоритм представляет собой адаптивную процедуру наискорейшего подъема для нахождения локальных мод плотности $f(x)$.

Процедура среднего сдвига порождает естественное разбиение множества X на классы: точки $x^{(i)}$ и $x^{(j)}$ принадлежат к одному классу, если итеративные процессы среднего сдвига, начинающиеся с этих точек, сходятся к одной и той же моде [2]. Эта процедура достаточно трудоемка, поэтому ее непосредственное применение ограничено выборками небольшого объема.

В работе [3] предложен алгоритм автоматической классификации, в котором процедура среднего сдвига применяется не для всех точек исходного множества X , а лишь для некоторого случайного подмножества из X . Этот прием позволяет увеличить объем обрабатываемых данных до нескольких десятков тысяч.

В типичных задачах тематической обработки аэрокосмических данных число классифицируемых объектов исчисляется сотнями тысяч и миллионами. Поэтому, использование этого алгоритма для обработки многозональных данных также приводит к неприемлемо большому вычислительным затратам.

В следующем разделе приводится описание быстрого алгоритма автоматической классификации многозональных данных, в котором стартовое множество точек для запуска процедуры среднего сдвига порождается клеточной структурой данных, организуемой в пространстве спектральных признаков.

Описание алгоритма

Предлагаемый алгоритм опирается на использование двух характерных особенностей многозональных данных. Первая из них заключается в ограниченности диапазонов изменения значений спектральных признаков (значения лежат в диапазоне целых чисел от 0 до $L - 1$, где L – число уровней квантования видеосигнала, обычно не превышающее 256), а вторая – в высокой частоте повторяемости векторов спектральных яркостей. Повторяемость обуславливается ограниченностью диапазона спектральных яркостей, наличием корреляции между спектральными диапазонами, а также относительной однородностью и достаточной протяженностью природных объектов. Алгоритм можно записать в виде следующей последовательности шагов.

1. *Формирование клеточной структуры данных в пространстве спектральных признаков.* Разбиваем все пространство значений спектральных признаков $[0, 255] \times \dots \times [0, 255]$ на гиперкубические клетки со стороной $2h$. Вводим общую нумерацию клеток (последовательно от одного слоя клеток к другому) и с каждой клеткой связываем набор попавших в нее спектральных векторов из X .

2. *Формирование таблицы частот повторяемости (“весов”) векторов множества X .* Эта таблица “весов” позволяет иногда в десятки раз сократить объем вычислений при определении выборочных средних и оценок плотностей распределения.

3. *Формирование множества начальных (стартовых) векторов S для запуска процедуры среднего сдвига.* Для каждой клетки, которая содержит более N_{\min} векторов из X , вычисляем вектор средних значений по всем точкам, попавшим в эту клетку. Совокупность полученных таким образом векторов образуют множество S .

4. *Оценка локальных мод плотности $f(x)$.* Применяем процедуру среднего сдвига, используя в качестве стартовых векторов элементы множества S . При этом стоит заметить, что для вычисления векторов $m_h(x)$ не нужно перебирать все элементы X , а достаточно использовать лишь элементы соседних к точке x клеток, которые легко определяются, благодаря введенной нумерации. В результате применения процедуры получаем множество локальных мод $Z_0 = \{z_1, \dots, z_{M_0}\}$.

5. *Формирование множества кандидатов в центры классов $Z_1 = \{z_1, \dots, z_{M_1}\}$.* Определяем на множестве Z_0 все подмножества близких друг к другу точек (точка близка к подмножеству, если она находится на расстоянии не большем h от некоторой точки этого подмножества). Для каждого выделенного подмножества вычисляем его вектор средних значений. Совокупность таких векторов и будет образовывать множество Z_1 .

6. *Нахождение множества центров классов $Z = \{z_1, \dots, z_M\}$.* Будем считать, что две точки из множества Z_1 принадлежат одному классу, если между этими точками нет “оврага” в плотности распределения $f(x)$. Существование “оврага” определяется для каждых двух точек z_i, z_j . Вдоль линии, соединяющей точки z_i и z_j (начиная с точки с меньшей плотностью), с шагом h вычисляем оценку плотности $\hat{f}_N(x)$. Если для некоторой точки x_m линии, соединяющей z_i и z_j , отношение между $\max[\hat{f}_N(x_1), \dots, \hat{f}_N(x_{m-1})]$,

где $\hat{f}_N(x_1) = \min[\hat{f}_N(z_i), \hat{f}_N(z_j)]$, и оценкой $\hat{f}_N(x_m)$ больше заданного порога T , то констатируем обнаружение “оврага”. В противном случае считаем, что точки z_i, z_j принадлежат одному классу. Точку (z_i или z_j) с меньшей плотностью удаляем из множества центров классов. Как показали эксперименты, в оценках $\hat{f}_N(x)$ вместо квадратичного ядра Епанечникова, целесообразно использовать мультипликативное треугольное ядро, потому что качество классификации при этом не ухудшается, а объем вычислений существенно сокращается. Так как ядро финитно, то при вычислении оценок $\hat{f}_N(x)$, как и при вычислении $m_h(x)$ в п. 3, достаточно использовать лишь элементы клеток, которые являются соседними к точке x .

7. *Распределение точек множества X по классам.* Используя классификатор ближайшего соседа, а в качестве обучающей выборки размеченное множество $Z_0 = \{z_1, \dots, z_{M_0}\}$, распределяем множество X по классам. Клеточная структура данных обеспечивает организацию быстрого поиска ближайшего соседа.

Результаты экспериментального исследования

Предложенный алгоритм программно реализован в среде Microsoft Visual Studio .NET и включен в пакет прикладных программ “GIPARD” [4], который предназначен для автоматизированного анализа данных дистанционного зондирования. Для работы алгоритма необходимо задать значения параметров: h , N_{\min} , T . Многочисленные экспериментальные исследования, проведенные как со спутниковыми данными, так и с аэроснимками, показывают, что классификацию многозонального изображения целесообразно проводить после осуществления линейного растяжения динамического диапазона значений яркости в каждой зоне на всю допустимую шкалу. Это преобразование значительно упрощает подбор оптимальных значений для параметра сглаживания h . В проведенных экспериментах наиболее качественные картосхемы получались при $h = 11$. Параметры N_{\min} и T оказывают слабое влияние на результат классификации. В экспериментах $N_{\min} \approx 50$, $T \approx 1.5$. Кроме этого, поскольку на реальных изображениях практически всегда присутствуют малопредставительные классы и, как правило, не представляющие интереса для пользователя, то при работе алгоритма предусмотрено объединение всех таких мелких классов в один класс-фон. Площадь класса-фона обычно составляет 0,5-1 % от общего изображения.

Ниже приводятся типичные результаты двух из многочисленного числа проведенных экспериментов. Обработка производилась на ПЭВМ Pentium IV с тактовой частотой 1.7 ГГц. Интерпретация картосхем производилась с использованием планов лесонасаждений и результатов визуально-инструментального дешифрирования спектрально-зональных аэроснимков. Для уменьшения раздробленности картосхем, получаемых в ходе попиксельной классификации многозональных изображений, целесообразно производить их постобработку медианным фильтром. Качество картосхем оценивалось специалистами-дешифровщиками и признано удовлетворительным.

Эксперимент 1. Обработывался фрагмент изображения (рис.1 (слева)) центральной части территории Красноярского края, полученного со спутника Landsat-7, размером 512×512 пикселей в 1, 2 и 3 каналах. Исходное число элементов изображения составляло 262144, параметры: $h = 12$, $N_{\min} = 50$, $T = 1.6$. Частота повторения пикселей – 74.5. Время обработки – 4 сек. На рисунке 1 (справа) представлена картосхема результатов классификации. На картосхеме представлено 11 значимых классов.

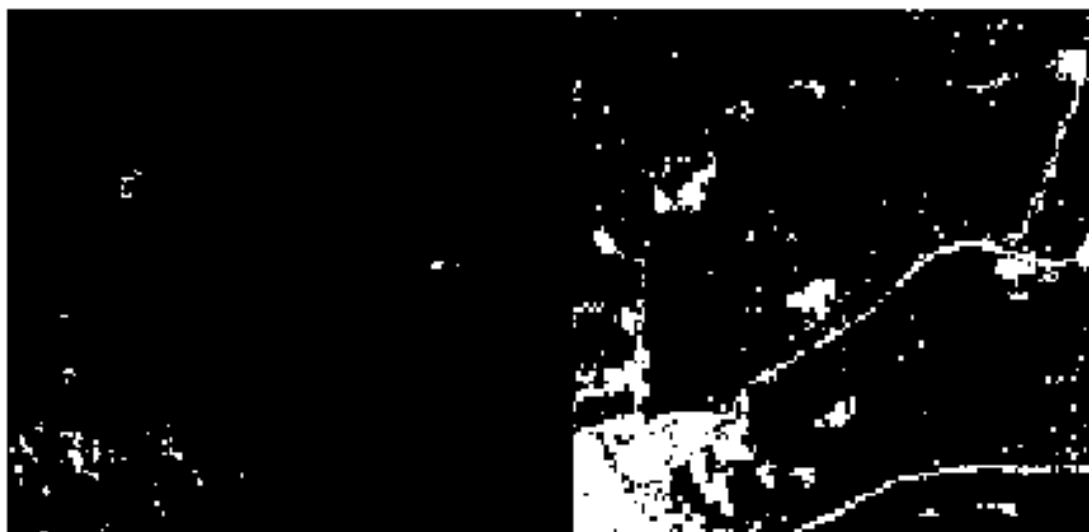


Рис. 1. Пример сегментации фрагмента изображения, полученного со спутника Landsat-7: слева – исходное изображение, справа – результат классификации.



Рис. 2. Пример сегментации фрагмента изображения, полученного с помощью самолетного сканера С-500: слева – исходное изображение, справа – результат классификации.

Эксперимент 2. Для обработки использовался фрагмент изображения (рис. 2 (слева)) Краснотуранского бора (юг Красноярского края), полученного с помощью самолетного сканера С-500 [5], размером 350×256 пикселей в 1, 2, 4 и 6 каналах. Исходное число элементов изображения составляло 89600, параметры: $h = 11$, $N_{\min} = 50$, $T = 1.5$. Обработка производилась с использованием трех и четырех каналов. При использовании трех каналов частота повторения значений пикселей составляла 4.7, время обработки – 5 сек. При обработке в четырех зонах спектра частота повторений – 1.35, время обработки – 18 сек. На рисунке 2 (справа) представлена картосхема результатов классификации.

Заключение

Представленный в докладе алгоритм классификации данных дистанционного зондирования не требует ни классифицированной обучающей выборки, ни каких-либо предположений относительно параметрической структуры данных, обеспечивая при этом высокое качество результатов. При проведении классификации от пользователя требуются минимальные усилия, связанные с настройкой параметров, для получения оптимальных результатов. Высокое быстродействие алгоритма обеспечивает возможность работы в диалоговом режиме. Предложенная процедура применима для обработки как космических, так и аэроснимков.

Список литературы

- [1] *Fukunaga K., Hostetler L.D.* The estimation of the gradient of a density function, with applications in patter recognition // IEEE Trans. Information Theory. - 1975. - Vol. 21, pp. 32-40.
- [2] *Cheng Y.* Mean shift, mode seeking, and clustering // IEEE Trans. Pattern Analysis and Machine Intelligence. - 1995. - Vol. 17, pp. 790-799.
- [3] *Comaniciu D., Meer P.* Distribution free decomposition of multivariate data // Pattern Analysis and Applications. - 1999. - No. 2, pp. 22-30.
- [4] *Снявский Ю.Н., Бужкина Е.А.* ППП “GIPARD” для автоматизированного анализа данных дистанционного зондирования // Матер. XLII Междунар. научн. студ. конф. “Студент и научно-технический прогресс”: Информационные технологии / Новосиб. гос. ун-т. - Новосибирск. - 2004. - С. 181-182.
- [5] *Аванесов Г.А., Баринов И.В., Глазков В.Д. и др.* Моделирование космического эксперимента с помощью самолетной лаборатории // Космические исследования земных ресурсов. М.: Наука, 1976. - С. 280-290.