RECOGNITION OF HYPERSPECTRAL IMAGES WITH USE OF CLUSTER ENSEMBLE AND SEMISUPERVISED LEARNING

Vladimir B. Berikov[1], *Igor A. Pestunov*[2], *Nikita M. Karaev*[3], *Ankit Tewari*[3,4]

[1]Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia
[2]Institute of Computational Technologies SB RAS, Novosibirsk, Russia
[3]Novosibirsk State University, Novosibirsk, Russia
[4]Birla Institute of Technology, Mesra Ranchi, Jharkhand, India

Abstract

We suggest a method for hyperspectral image analysis on the basis of semi-supervised learning. The main idea is to divide the process of training of a classifier into two stages. First of all, with usage of cluster ensemble algorithms, variants of image segmentation are obtained. On their basis, the averaged co-association matrix is calculated. On the second stage, a classifier is constructed on labeled pixels using similarity based learning algorithms with the given matrix as input. An example of the application of the method for analysis of hyperspectral images is given. It is shown that the suggested algorithm is more robust to noise than the standard support vector machine method.

Keywords: cluster ensemble, learning by similarity, semi-supervised learning, hyperspectral image

РАСПОЗНАВАНИЕ ГИПЕРСПЕКТРАЛЬНЫХ ИЗОБРАЖЕНИЙ С ИСПОЛЬЗОВАНИЕМ КЛАСТЕРНОГО АНСАМБЛЯ И ЧАСТИЧНО КОНТРОЛИРУЕМОГО ОБУЧЕНИЯ

Бериков В.Б.[(1),(2)], Пестунов И.А.[(3)], Караев Н.М.[(2)], Тевари А.[(2),(4)]
[1] Институт математики им. С.Л. Соболева СО РАН, г. Новосибирск
[2] Новосибирский государственный университет, г. Новосибирск
[3] Институт вычислительных технологий СО РАН, г. Новосибирск
[4] Birla Institute of Technology, Mesra Ranchi, Jharkhand, India

Предлагается метод анализа гиперспектральных изображений на основе частично контролируемого обучения. Основная идея состоит в разделении процесса обучения на два этапа. Вначале с помощью ансамбля алгоритмов кластерного анализа строятся варианты сегментации изображения. Далее вычисляется усредненная коассоциативная матрица. На втором этапе по размеченным пикселям строится решающая функция с применением алгоритмов обучения по сходству, на вход которого подается полученная матрица. Описан пример применения разработанного метода для анализа гиперспектральных изображений. Показано, что предложенный алгоритм более устойчив к шуму, чем стандартный метод опорных векторов.

Ключевые слова: кластерный ансамбль, обучение по сходству, частично контролируемое обучение, гиперспектральное изображение.

Введение. В дистанционном зондировании Земли активно используются средства и технологии гиперспектральной съемки в видимом и ближнем инфракрасном диапазонах спектра [1]. Особенностями гиперспектральных изображений являются большое число спектральных каналов (до нескольких сотен) и малая спектральная ширина каждого канала (порядка нескольких нанометров). Одной из важных задач, возникающих при анализе таких изображений, является их распознавание, т.е. отнесение отдельных пикселей или участков изображения к одному из классов. При этом задание обучающей информации о классах - достаточно трудоемкая операция, требующая ручной разметки и слабо поддающаяся автоматизации. Для построения классификатора в случае, когда лишь для сравнительно малой части выборки известны метки классов, применяются методы частично контролируемого обучения (полу-обучения, semi-supervised learning). Существует достаточно большое число таких методов [2,3], однако их применение при анализе гиперспектральных изображений затрудняется отсутствием сведений о вероятностной структуре данных, их большой размерностью, наличием зашумленных каналов [4,5].

Цель данной работы заключается в разработке нового алгоритма для решения задачи полуконтролируемого обучения для сложноструктурированных, зашумленных данных большого объема; его теоретическом обосновании и применении при анализе гиперспектральных изображений. Новизна работы состоит в сочетании алгоритмов коллективного кластерного анализа [6] и методов обучения по сходству [7].

Идея предлагаемого алгоритма состоит в следующем. Процесс построения решающей функции делится на два основных этапа. На первом этапе с помощью набора различных алгоритмов кластерного анализа по таблице данных строятся варианты группировки множества объектов, определяются индексы качества. На основе полученных вариантов вычисляется усредненная с весами коассоциативная матрица (co-association matrix). Элементы матрицы равны средневзвешенной частоте отнесения пар объектов к одинаковым кластерам по всем вариантам разбиения, а веса зависят от оценочных функций (индексов качества, мер разнообразия вариантов). В определенном смысле, матрица задает меры похожести объектов в новом признаковом пространстве, полученном из исходного с помощью некоторого неявного преобразования.

На втором этапе строится решающая функция с применением алгоритма, на вход которого поступает сформированная коассоциативная матрица. В качестве такого алгоритма может выступать любой из известных, основанных на обучении по сходству или использовании

ядра (kernel based) [7], например, метод опорных векторов (Support Vector Machine), ядерный дискриминант Фишера (Kernel Fisher Discriminant), ядерная версия алгоритма ближайших соседей (Kernel kNN).

Привлечение коллектива алгоритмов кластерного анализа позволяет повысить устойчивость решений, более точно восстановить метрические отношения между объектами в условиях шумовых искажений и наличия сложных структур данных, что в конечном итоге повышает качество распознавания. В качестве базовых алгоритмов на этапе построения коллективного группировочного решения используются алгоритмы, имеющие линейную трудоемкость (например, алгоритм К-средних).

Постановка задачи полуконтролируемого обучения. Пусть имеется генеральная совокупность объектов распознавания X и конечное множество меток классов Y. Все объекты описываются числовыми признаками.

При заданных признаках $f_1,..,f_m$ вектор $x=(f_1(x),..,f_m(x))$ называется признаковым описанием объекта $x \in X$. Далее мы отождествляем объект и его признаковое описание. В задаче полуконтролируемого обучения на вход подается выборка $X_N = \{x_1,..,x_N\}$ объектов из X. В этой выборке присутствуют объекты двух типов: $X_c = \{x_1,..,x_k\}$ - размеченные объекты с заданными классами, которым они принадлежат: $Y_c = \{y_1,..,y_k\}$; $X_u = \{x_{k+1},..,x_N\}$ - неразмеченные объекты.

В различных вариантах постановки задачи требуется либо провести т.н. индуктивное обучение - построить алгоритм классификации $a: X \to Y$, который будет, минимизируя вероятность ошибки, сопоставлять классы объектам их X_u, а также новым объектам X_{test}, которые были недоступны на момент построения алгоритма, либо требуется провести трансдуктивное обучение - получить метки классов только для объектов из X_u с минимальной ошибкой. В данной работе рассматривается второй вариант постановки задачи.

Коллективные решения в кластерном анализе. Задачей кластерного анализа является разбиение выборки на непересекающиеся подмножества, называемые кластерами, так чтобы каждый кластер представлял группу похожих объектов, а объекты в разных кластерах существенно различались. В настоящее время в кластерном анализе широко применяется коллективный подход, который позволяет получать более устойчивые группировочные решения. Существует несколько вариантов получения коллективного решения задачи кластерного анализа: использование т.н. матрицы усредненного попарного сходства, максимизация степени согласованности решений (с помощью исправленного индекса Ранда, нормализованной взаимной информации и т.д.), применение теоретико-графовых методов. В предлагаемом в данной работе алгоритме используется матрица усредненного попарного сходства. Для построения матрицы кластеризация всех поданных на вход объектов X коллективом различных алгоритмов $\mu_1,..,\mu_M$ кластерного анализа. Каждый алгоритм дает L_m вариантов разбиения, $m=1,..,M$. По результатам работы алгоритмов составляется матрица H усредненных попарных различий объектов из X. Элементы матрицы равны:

$$h(i,j)=\sum_{m=1}^{M}\alpha_m\frac{1}{L_m}\sum_{l=1}^{L_m}h_{lm}(i,j), \qquad (1)$$

где $i,j \in \{1,..,N\}$ - номера объектов $(i \neq j)$, $\alpha_m \geq 0$ - заданные веса такие, что $\sum_{m=1}^{M}\alpha_m=1$, $h_{lm}(i,j)=0$, если пара (i,j) принадлежит разным кластерам в l-ом варианте разби-

ения, полученного алгоритмом μ_m и 1, если принадлежит одному кластеру. Способ нахождения оптимальных весов, минимизирующих оценку погрешности классификации был предложен в работе [8].

Ядерные методы классификации. Для решения задачи классификации с учителем широко распространены ядерные методы, в основе которых лежит понятие ядра (kernel). Подбор ядра определяет переход в «спрямляющее» пространство и позволяет применять линейные алгоритмы классификации к линейно неразделимой выборке [7].

В ядерных методах классификации широко известна теорема Мерсера [9], которая устанавливает необходимое и достаточное условие на то, чтобы функция была ядром:

Теорема (Мерсер). Функция $K(x,x')$ является ядром тогда и только тогда, когда она симметрична, $K(x,x')=K(x',x)$, и неотрицательно определена: для любой конечной выборки $X^p=(x_1,...,x_p)$ из X матрица $K=\|K(x_i,x_j)\|$ размера $p\times p$ неотрицательно определена: $z^TKz\geq 0$ для любого $z\in\mathbb{R}^p$.

Идея алгоритма состоит в построении матрицы похожести (1) для всех объектов из подаваемой на вход выборки X: чем чаще пара объектов попадает в один и тот же кластер, тем более похожими друг на друга мы их будем считать. Нами доказано следующее

Утверждение. Функция (1) удовлетворяет условиям теоремы Мерсера.

Таким образом, функция H может быть использована в ядерных методах классификации, в частности, в методе опорных векторов SVM.

Алгоритм CASVM. Ниже описаны шаги алгоритма полуконтролируемого обучения, сочетающего ансамблевый кластерный анализ и метод опорных векторов.

Вход: объекты X_c с заданными классами Y_c и объекты X_u, число алгоритмов кластеризации M, число кластеризаций L_m каждым алгоритмом $\mu_m, m=1,...,M$.

Выход: классы объектов X_u.

1. Провести кластеризацию объектов $X_c\cup X_u$ алгоритмами $\mu_1,..,\mu_M$ кластерного анализа, получив L_m вариантов разбиения от каждого алгоритма $\mu_m, m=1,...,M$.
2. Вычислить матрицу H на $X_c\cup X_u$ по формуле (1).
3. Обучить SVM на размеченных данных X_c, используя матрицу H в качестве ядра.
4. С помощью SVM предсказать классы для неразмеченных объектов X_u.

Конец алгоритма.

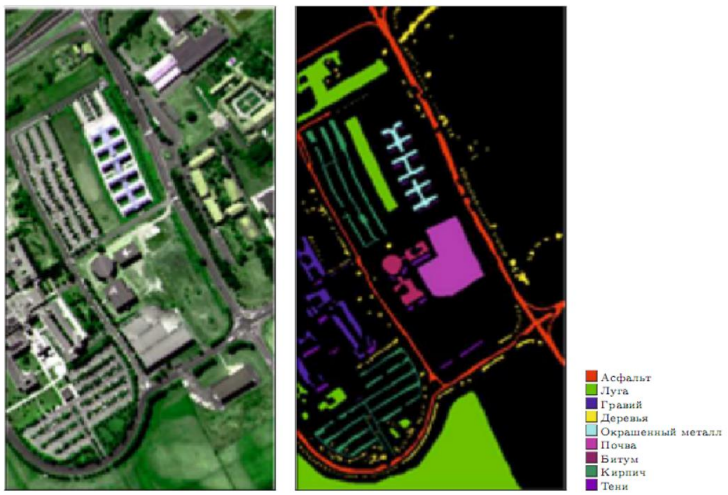
Отметим, что в предложенном алгоритме не требуется хранить в памяти матрицу H размера $N\times N$ целиком: достаточно хранить матрицу кластеризаций размера $N\times L$, где $L=\sum_{l=1}^{M}L_m$, в этом случае матрицу H можно вычислять динамически. В прикладных задачах как правило $L<<N$, например, при работе с пикселями изображений.

Анализ гиперспектрального изображения. Для экспериментального исследования алгоритма был проведен эксперимент с изображением Pavia University scene размером 610 на 340 пикселей, которое содержит 103 спектральных канала. Пространственное разрешение снимка составляет 1.3 м. На рисунке 1а) показан RGB-композит изображения (каналы 40, 50 и 70), а на рисунке 1б) приведено эталонное разбиение изображения на тематические классы.

Отметим, что на снимке имеются неразмеченные пиксели, которые не отнесены ни к одному из девяти классов. Данные пиксели были исключены из рассмотрения при анализе.

При экспериментальном исследовании алгоритма 1% пикселей, отобранных случайным образом для каждого класса, составили размеченную выборку; оставшиеся были включены в неразмеченную часть. Для изучения влияния шума на качество работы алгоритма, случайно

отобранные r% значений спектральных яркостей пикселей в разных каналах подвергались искажающему воздействию: соответствующее значение x заменялось величиной, выбранной случайным образом из интервала $[x(1-p), x(1+p)]$, где r, p - заданные параметры. Зашумленная таблица данных, содержащая значения спектральных яркостей пикселей по всем каналам, подавалась на вход алгоритма CASVM, а котором в качестве базового алгоритма для построения кластерного ансамбля был выбран алгоритм K-средних. Различные варианты разбиения получались варьированием числа кластеров в интервале $[30, 30+L]$, где L было равно 120. Кроме того, для построения каждого варианта решения случайным образом выбирались каналы, число которых было задано двум. Для ускорения работы алгоритма K-средних и получения более разнообразных вариантов группировки, число его итераций было ограничено значением 1.



а б

Рис. 1. Гиперспектральное изображение Pavia University scene (RGB композит) (а) и размеченные данные (б).

Поскольку предложенный алгоритм реализует идею обучения метрике расстояния (distance metric learning), было бы естественно провести его сравнение с аналогичным алгоритмом (нашем случае - методом опорных векторов SVM), использующим стандартную евклидову метрику, в аналогичных условиях (выбирались параметры алгоритма, рекомендуемые по умолчанию в среде Матлаб). В таблице показаны значения точности классификации неразмеченных пикселей изображения Pavia University scene для некоторых значений параметров зашумленности. Время работы алгоритма составило около 2 мин на двухъядерном процессоре Intel Core i5 с тактовой частотой 2.8 ГГц и объемом оперативной памяти 4 Гбайт. Как видно из таблицы, алгоритм CASVM обладает большей устойчивостью к шуму, чем алгоритм SVM.

Точность алгоритмов CASVM и SVM при различных значениях параметров шума.

Параметры шума r, p	0%, 0	10%, 0.1	20%, 0.2	30%, 0.3
CASVM	0.82	0.80	0.78	0.77
SVM	0.83	0.75	0.66	0.64

Заключение. В работе рассмотрен один из вариантов постановки задачи распознавания образов – задача полуконтролируемого обучения. Был разработан алгоритм CASVM для решения этой задачи. Он основывается на сочетании методов коллективного кластерного анализа и ядерных методов классификации. Проведено экспериментальное исследование предложенного алгоритма на гиперспектральном изображении. Показано, что алгоритм CASVM более устойчив к шуму, чем стандартный метод опорных векторов SVM.

ЛИТЕРАТУРА

[1] Бондур В.Г. Современные подходы к обработке больших потоков гиперспектральной и многоспектральной аэрокосмической информации // Исследование Земли из космоса. 2014. N 1. С. 4-16.

[2] Zhu X. Semi-supervised learning literature survey / Tech. Rep. (Department of Computer Science, Univ. of Wisconsin, Madison, 2008), no. 1530.

[3] Wang, F. Label propagation through linear neighborhoods / Wang, F., Zhang, C. // ICML06, 23rd International Conference on Machine Learning. Pittsburgh, USA.

[4] Куликова Е.А., Пестунов И.А. Классификация с полуобучением в задачах обработки многоспектральных изображений // Вестник Казахского национального университета. Серия: Математика, механика, информатика. 2008. Т. 13. № 3. С. 284-290.

[5] Wang L., Hao S., Wang Q., Wang Y. Semi-supervised classification for hyperspectral imagery based on spatial-spectral Label Propagation // ISPRS Journal of Photogrammetry and Remote Sensing. 2014. Vol. 97. P. 123-137.

[6] Berikov V., Pestunov I. Ensemble clustering based on weighted co-association matrices: Error bound and convergence properties // Pattern Recognition. 2017. Vol. 63. P. 427-436.

[7] Shawe-Taylor J., Cristianini N. Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.

[8] Berikov V.B. Weighted ensemble of algorithms for complex data clustering // Pattern Recognition Letters. 2014. Vol. 38. P. 99-106.

[9] Mercer J. Functions of positive and negative type and their connection with the theory of integral equations / Philos. Trans. Roy. Soc. London. 1909.