Алгоритм управляемой классификации изображений дистанционного зондирования Земли с использованием иерархических гистограмм

А.Ю. Денисова[a], В.В. Сергеев[a]

[a] *Самарский национальный исследовательский университет, 443086, Московское шоссе, 34, Самара, Россия*

Аннотация

Работа посвящена применению для управляемой классификации структуры гистограммы-дерева, предложенной авторами для построения гистограмм многоканальных изображений. Алгоритм классификации строится путём модификации структуры данных иерархической гистограммы для реализации простых правил классификации, получаемых по обучающей выборке. Предложенный алгоритм исследуется на наборе гиперспектральных изображений. В статье производится сравнение предлагаемого алгоритма с известным классификатором на основе дерева решений – алгоритмом C4.5.

Ключевые слова: управляемая классификация, многоканальные изображения, иерархические гистограммы

1. **Введение**

В обработке изображений широкое распространение получила такая кусочно-постоянная оценка плотности распределения, как гистограмма. Гистограммы применяются для цветовой коррекции, бинарной обработки, кластеризации изображений и, конечно же, классификации. В задачах классификации гистограмма выступает, как правило, либо к качестве самостоятельного признака объекта (локальная гистограмма), либо в качестве эмпирической оценки плотности распределения признаков (глобальная гистограмма).

Множество алгоритмов классификации в зависимости от способов применения гистограммы изображения можно подразделить на следующие группы:

1) классификаторы, использующие оценку плотности для каждого класса. К данной категории относятся управляемые алгоритмы, получающие свою гистограмму для каждого класса по обучающей выборке. Полученные многомерные распределения векторов признаков позволяют использовать статистические алгоритмы классификации для принятия решения, например, байесовский классификатор или классификатор максимального правдоподобия [1].

2) классификаторы, использующие модель смеси распределений для описания плотности вероятностей всей совокупности данных. Данный подход следует отнести к неуправляемым алгоритмам классификации, так как после оценки глобальной плотности распределения признаков производится разделение смеси распределений, эквивалентное определению количества классов и их эмпирических распределений.

Непосредственное использование глобальной гистограммы ограничено малыми размерностями данных (от двух до четырёх компонент в зависимости от количества бит, используемых для представления данных в одном канале изображения [3]), из-за высокой вычислительной сложности и требований к объему памяти в многомерном случае. Для устранения этого ограничения существует несколько подходов, основанных на использовании в алгоритмах классификации иерархических структур данных, а именно деревьев. Дерево может быть использовано как для аппроксимации и хранения гистограммы (алгоритм TUBE (Tree-based Unsupervised Bin Estimator) [4], алгоритм DET (Density Estimation Trees) [5], алгоритмы «в глубину» и «в ширину» предложенные авторами в работе [3] и другие), так и для хранения последовательности решающих правил, позволяющих произвести классификацию. В первом случае оптимизируется только время построения многомерной гистограммы, дальнейшая процедура классификации реализуется некоторым известным алгоритмом. Во втором случае каждый узел дерева сохраняет простое решающее правило, которое разделяет пространство признаков на области в соответствии с требуемыми целевыми параметрами классификации.

Алгоритмы на основе иерархического разбиения признакового пространства используют обучающую выборку для построения решающего правила на каждом уровне иерархии. Основными этапами построения классификатора являются: выбор признака, по которому производится деление признакового пространства, вычисление границы разбиения, проверка условия остановки построения дерева. Условие остановки необходимо для решения проблемы переобучения и сокращения вычислительной сложности процедуры обучения. Поскольку решающие правила данных алгоритмов образуют иерархическое разбиение пространства признаков, а способы их вычисления основаны на статистических характеристиках классов в обучающей выборке в каждой из ячеек подпространства, то можно считать, что данные алгоритмы неявно производят оценку плотности распределения векторов признаков.

Предлагаемый в данной работе алгоритм основан на модификации гистограммы-дерева, предложенной в статье [3]. В отличие от других алгоритмов классификации он ориентирован на использование естественной группировки данных в пространстве признаков в соответствии с битовым представлением вектора признаков. Это делает возможным быстрое построение дерева за счет использования бинарных операций. В нашем случае структура данных имеет фиксированную максимальную глубину и не является бинарной. Дополнительным свойством предложенного алгоритма

является возможность извлечения из получаемой структуры данных гистограммы изображения и классов, рассчитанных на этапе обучения классификатора.

Подробное описание модифицированной структуры данных, алгоритма классификации и результатов экспериментальных исследований приведено в последующих разделах статьи. Экспериментальные исследования содержат сравнение качества работы предложенного алгоритма с алгоритмом C4.5, который реализует подход «дерево решений» и был отмечен в [1] как один из лучших классификаторов.

2. Алгоритм классификации с использованием гистограммы-дерева

2.1. Постановка задачи

Обозначим за $x_n = (x_{n0},..., x_{nL-1}), n = \overline{1,..., N}$ – вектор признаков n-ого пикселя изображения и будем полагать, что $x_{ns} \in [0, 2^B - 1], s = 0,..., L-1$, где B – разрядность двоичного представления компонент вектора признаков, L – размерность пространства признаков, N – общее количество классифицируемых векторов. Пусть все пиксели изображения относятся к одному из C классов $\Omega_i, i = 0,..., C-1$. Задача классификации состоит в установлении соответствия между значением вектора признаков x_n и номером класса $\omega_n \in \{0,..., C-1\}$.

Будем рассматривать управляемую классификацию, когда решающее правило строится на основании некоторой обучающей выборки векторов признаков $y_k = (y_{k0},..., y_{kL-1}),\ k = 1,..., T$, где T – объём выборки, для которой известны метки классов $\omega_k \in \{0,..., C-1\}$.

2.2. Структура данных

В настоящей работе для реализации алгоритма классификации была модифицирована структура данных, предложенная в [3], так называемая «гистограмма-дерево». Каждый уровень $1 \le r \le B$ гистограммы-дерева, кроме начального нулевого $r = 0$, содержащего один корневой элемент, определяет разбиение пространства на ячейки-гиперкубы со стороной $2^{(B-r)}$. Пример разбиения пространства для двумерного случая при $B = 8$ представлен на рис. 1.

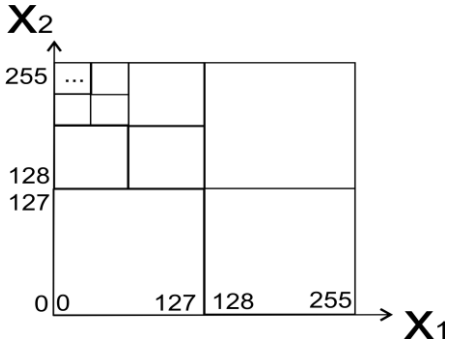

Рис. 1. Пример разбиения, соответствующий гистограмме-дереву в случае $L = 2$ и $B = 8$.

Будем далее называть маской узла число $q^{(r)}$, сформированное из r-тых бит в компонент вектора признаков. Здесь и ниже будем полагать, что нулевая битовая плоскость соответствует старшему разряду в битовом представлении вектора признаков, а B-ая плоскость младшему разряду.

Маска узла является номером ветви, к которой относится значение x_n на данном уровне r. Последовательность масок узлов от первого уровня до последнего $q^{(1)}, q^{(2)},..., q^{(B)}$ представляет собой путь к терминальному узлу, однозначно определяющему значение всех компонент отсчёта x_n. Для построения гистограммы-дерева достаточно хранить в узле дерева маску узла $q^{(r)}$ и «частоту узла» f, которая равна числу векторов признаков, старшие r бит которых определяются масками $q^{(1)}, q^{(2)},..., q^{(r)}$ узлов на пути к рассматриваемому узлу от корневого. Описанная структура представляет собой несбалансированное дерево глубины B. Максимальное число потомков каждого узла составляет 2^L, однако на практике, имеет смысл создавать лишь узлы с не нулевым значением частоты f. Терминальные узлы дерева соответствуют конкретному значению отсчёта.

Применение к задаче классификации описанной выше гистограммы-дерева предполагает необходимость модификации дерева для хранения информации о распределении объектов выборки в каждый класс. Для этого каждый узел дерева пополняется массивом чисел $f_0, f_1,..., f_{C-1}$, соответствующим «частоте» попадания объектов заданного класса в ячейку, определяемую узлом дерева.

2.3. Алгоритм классификации

Обучение предлагаемого алгоритма классификации заключается в построении модифицированной гистограммы-дерева. Для построения дерева можно воспользоваться алгоритмами «в глубину» и «в ширину», предложенными и описанными в [3].

Пусть $y_k = (y_{k0},..., y_{kL-1}),\ k = 1,..., T$ – обучающая выборка, которой соответствуют метки классов $\omega_k \in \{0,..., C-1\}$.

Алгоритм «в глубину» выполняет рекурсивное построение дерева последовательно для каждого отсчёта, начиная со старшего битового уровня и заканчивая младшим. При старте алгоритма корневой узел на уровне $r = 0$ устанавливается текущим. На каждом уровне r рекурсии для текущего узла значение частоты узла f и частота f_i пикселей, принадлежащих узлу классу Ω_i увеличивается на единицу. Затем для текущего узла из вектора y_n извлекается битовая плоскость $r+1$, соответствующая маске $q^{(r+1)}$ и уровню дерева. Если среди потомков текущего узла нет дочерних улов с маской $q^{(r+1)}$, то добавляется новый узел, который становится текущим, иначе текущим узлом становится узел с маской $q^{(r+1)}$. Счётчик текущего уровня дерева становится равным $r = r+1$. Обработка вектора y_n заканчивается при достижении значения $r = B$, при этом список дочерних узлов текущего узла устанавливается пустым и узел становится терминальным.

Второй алгоритм построения дерева – алгоритм «в ширину». Алгоритм выполняет построение дерева послойно, для построения каждого слоя необходим один проход по изображению. Заполнение текущего уровня происходит аналогично предыдущему алгоритму с той лишь разницей, что новые потомки могут быть созданы только в рамках текущего заполняемого уровня дерева.

Полученное в результате обучения дерево на каждом уровне $r \le B$ представляет собой гистограмму изображения с длиной интервала 2^{B-r} по каждой компоненте вектора признаков. Список частот $f_0, f_1,..., f_{C-1}$ попадания пикселей обучающей выборки в классы $\Omega_i, i \in \{0,1,..., C-1\}$, хранимый в каждом узле дерева, соответствует вероятностям попадания пикселей, относящихся к ячейке гистограммы, задаваемой узлом.

Таким образом, для реализации процедуры классификации достаточно отыскать интервал гистограммы, которому соответствует заданный вектор признаков x_n и применить следующее решающее правило: пиксель относится к тому классу, частота попадания в который для заданного интервала гистограммы (узла дерева) максимальна. Используемое в данной работе решающее правило имеет вид:

$$x_n \in \Omega_j,\ \textit{если}\ \ f_j = \max_{i=0,..., C-1} f_i\ .$$

Терминальные узлы для дерева глубины B однозначно соответствуют отсчётам гистограммы изображения с единичным интервалом по каждой компоненте вектора-пикселя, если использовать в качестве частоты значение f. Если в качестве значения частоты брать f_i, то последовательность терминальных узлов задаёт распределение векторов признаков в заданный класс. Таким образом, полученное дерево описывает как гистограмму совместного распределения признаков, так и распределения признаков по классам. Очевидно, что построение полного дерева (глубины B) соответствует ситуации переобучения классификатора и приводит к тому, что классифицироваться будут только вектора признаков совпадающие с векторами обучающей выборки. Из чего следует необходимость построения гистограммы с более широким интервалом.

Используемая структура данных позволяет производить аппроксимацию гистограммы путём отбрасывания узлов с малой частотой f, что эквивалентно представлению гистограммы с неравномерным размером ячейки. При этом ячейки большего размера соответствуют менее вероятным значениям векторов признаков.

Обозначим пороговую частоту $\nu_{\min}$. Алгоритмы построения дерева для учёта ограничения на минимальную частоту узла $\nu_{\min}$ модифицируются следующим образом: в алгоритме «в глубину» после построения полного дерева, начиная с терминальных и продвигаясь к корневому узлу, удаляются все узлы, частота f которых меньше чем $\nu_{\min}$. В алгоритме «в ширину» узлы с частотой $f < \nu_{\min}$ не подвергаются дальнейшему делению. Оба алгоритма дают идентичные структуры данных на выходе, однако алгоритм «в ширину» является более эффективным с точки зрения затрат памяти во время выполнения обучения.

Значение пороговой частоты не должно превышать минимальной мощности множества векторов признаков из обучающей выборки, относимых к одному классу. Выбор значения $\nu_{\min}$ может быть сделан из соображений баланса между требованиями экономии памяти и требуемой точности классификации.

При классификации с использованием усеченного дерева, возможна ситуация, когда заданный вектор признаков не попадёт ни в один из интервалов гистограммы, в силу конечности обучающей выборки. Будем считать, что такие вектора относятся в дополнительный класс Ω_C, соответствующий ситуации неопределённости решения.

Предлагаемый классификатор позволяет организовать быстрое вычисление гистограмм векторов признаков по классам, реализуя при этом простейшее правило классификации по максимуму условной вероятности попадания в класс. Следует отметить, что работа классификатора не требует от данных соответствия гауссову или какому-либо другому закону распределения в силу использования непараметрической оценки плотности вероятностей.

3. Результаты экспериментальных исследований

Для экспериментальных исследований качества работы предлагаемого алгоритма классификации был использован набор гиперспектральных изображений дистанционного зондирования Земли «Hyperspectral Remote Sensing Scenes» [6], полученных с помощью сенсоров AVIRIS, ROSIS и HYPERION. Данные изображения имеют разметку классов объектов, составленную данным по наземных наблюдений. Каждому пикселю изображения, принадлежащему разметке, соответствует определённый индекс класса или нулевой индекс, если информация о классе отсутствует. Список использованных изображений и характеристики классов для них приведёны в таблице 1.

Таблица 1. Тестовые изображения и их характеристики

Название	Сенсор	Размер, пикс.	Количество каналов	Количество классов	% размеченных пикселей от общего числа пикселей	Количество размеченных пикселей, пикс.	Минимальная мощность класса, пикс.	Максимальная мощность класса, пикс.	Средняя мощность класса, пикс.
Pavia	ROSIS	1096×715	102	9	18.91	148186	2685	65971	16461.33
Pavia University	ROSIS	610×340	103	9	20.63	42787	947	18649	4752.889
Salinas	AVIRIS	512×217	220	16	48.72	54130	916	11271	3383.063
Kennedy Space Center	AVIRIS	512×614	176	13	1.66	5219	105	927	400.8462
Botswana	Hyperion	1476×256	145	14	0.86	3250	95	314	232
Indian pines	AVIRIS	145×145	220	16	48.75	10250	20	2455	640.5625

В качестве векторов признаков применялись первые коэффициенты разложения изображений методом главных компонент. Поскольку предложенный классификатор работает с целочисленными векторами признаков, то множество векторов признаков предварительно преобразовывалось с помощью линейного контрастирования и равномерного квантования в изображение с $B=8$ битами на отсчёт в каждом канале. Для случая трёхмерных векторов признаков изображения признаков и разметка на классы приведены на рис. 2-3.

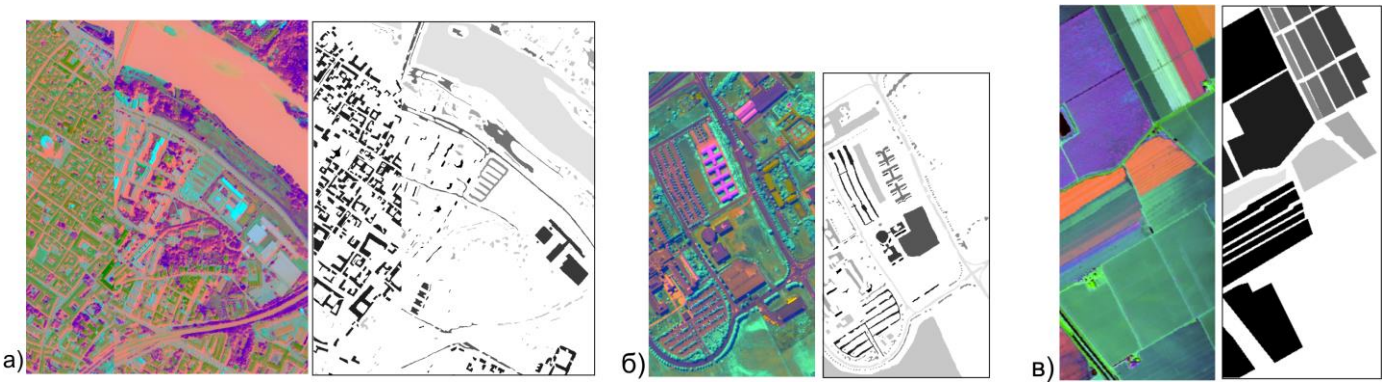

Рис. 2. Тестовые гиперспектральные изображения и разметка на классы: а) Pavia, б) Pavia University, в) Salinas. Белый цвет разметки соответствует отсутствию информации о классе.

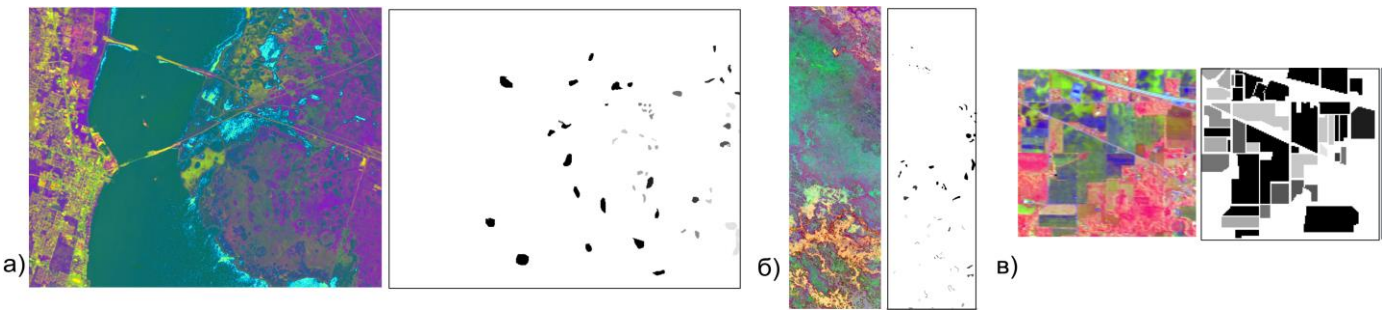

Рис. 3. Тестовые гиперспектральные изображения и разметка на классы: а) Kennedy Space Cente, б) Botswana, в) Indian pines. Белый цвет разметки соответствует отсутствию информации о классе.

Обучающая и контрольная выборки формировались M раз для каждого тестового изображения случайным образом. При этом, все множество размеченных на классы пикселей делилось между обучающей и контрольной выборкой в соотношении 75:25 соответственно. Для каждой из M пар пиксели в обучающей и контрольной выборке не пересекались. Пример разметки классов, обучающей и контрольной выборок для одного из изображений представлен на рис. 4.

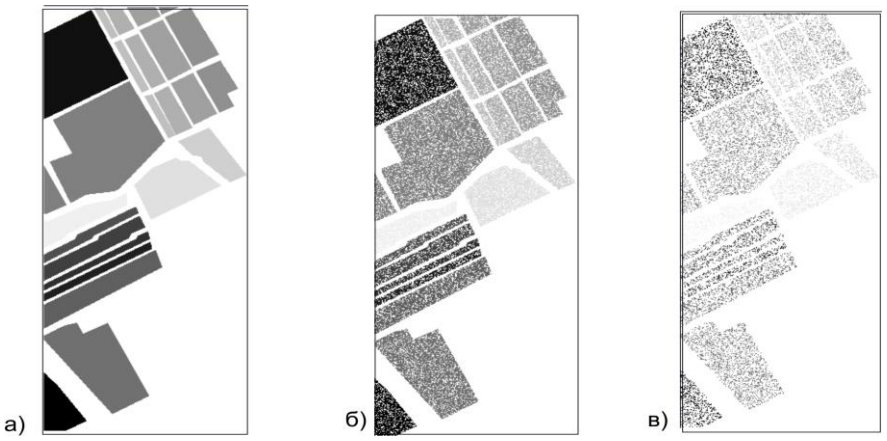

Рис. 4. Разметка классов: а) исходная, б) тестовая, в) контрольная. Белый цвет соответствует отсутствию информации о классе.

Оценка качества работы алгоритмов использовалась с использованием средней вероятности верной классификации p по результатам $M = 10$ запусков алгоритма обучения и классификации:

$$p = \sum_{i=1}^{M} \frac{\eta_i}{K},$$

где η_i – количество правильно проклассифицированных отсчётов контрольной выборки для i-ого запуска алгоритма, K – общее количество отсчётов контрольной выборки.

Ниже приведены результаты экспериментов для алгоритма на основе гистограммы-дерева и алгоритма C4.5 [7]. Для построения модифицированной гистограммы-дерева использовался алгоритм «в ширину», позволяющий сокращать издержки памяти во время выполнения программы. В качестве алгоритма C4.5 была использована его реализация из открытого пакета программ Accord .Net Framework [8] с параметрами, взятыми по умолчанию.

Исследование минимальной частоты узла ν_{min} для алгоритма «в ширину» было произведено на изображении Salinas. В таблице 2 приведена *средняя вероятность верной классификации по контрольной выборке* предложенным алгоритмом изображения Salinas при различных значениях ν_{min}.

Таблица 2. Средняя вероятность верной классификации по контрольной выборке изображения Salinas предложенным алгоритмом

ν_{min}	Размерность признакового пространства			
	3	6	9	12
5	0.78	0.66	0.62	0.60
15	0.86	0.80	0.75	0.72
25	0.87	0.84	0.79	0.77
35	0.88	0.85	0.81	0.79

Из таблицы 2 видно, что наибольшая вероятность верной классификации соответствует случаю $\nu_{min} = 35$ для всех рассматриваемых значений размерности вектора признаков. Большие значения частоты (в эксперименте были исследованы также значения $30 < \nu_{min} \le 100$) не приводили к существенному увеличению вероятности верной классификации и колебались относительно приведённого в таблице значения $\nu_{min} = 35$. Таким образом, значение частоты узла $\nu_{min} = 35$ соответствовало окончанию участка устойчивого роста вероятности верной классификации. Значение параметра $\nu_{min} = 35$ было использовано во всех последующих экспериментах с остальными изображениями набора.

Средние значения вероятности верной классификации на контрольной выборке для предложенного алгоритма с параметром $\nu_{min} = 35$ и для алгоритма C4.5 приведены в таблице 3.

Как видно из таблицы 3, с увеличением размерности векторов признаков вероятность верной классификации для алгоритма с гистограммой-деревом убывает. Так как размерность вектора наращивается за счет все менее информативных коэффициентов разложения методом главных компонент, то получаемое в пространстве более высокой размерности облако точек менее плотно и получаемая аппроксимация гистограммы менее точная. В случае использования трёх главных компонент разложения, предложенный классификатор обеспечивает высокие вероятности

верной классификации для четырех из шести изображений набора: Pavia, Pavia University, Salinas и Botswana. Причина менее удачной классификации изображения Indian pines заключается в наименьшем объёме обучающей выборки по сравнению с другими изображениями (см. таблицу 1), а также в том, что минимальная мощность класса в его случае составляла всего 20 пикселей, что меньше ограничения на частоту узла $\nu_{min} = 35$ и в процессе обучения пиксели данного класса могли войти в состав другого более крупного класса. Изображение Kennedy Space Center характеризовалось наличием «битых пикселей» в каналах, то есть пикселей с аномально завышенными значениями яркости. В такой ситуации, при реальном динамическом диапазоне значений до 1244 на исходном изображении, наличие пикселей со значением яркости близкими к 65535 привело к тому, что ряд каналов данного изображения имел завышенные значения дисперсии, и это оказало влияние на коэффициенты разложения методом главных компонент. Поскольку «битые» пиксели наблюдались в случайных позициях в различных каналах, то отнести их к дополнительному классу или исключить из обучающей и контрольной выборок не представлялось возможным.

Таблица 3. Средняя вероятность верной классификации по контрольной выборке исследуемыми алгоритмами

Изображение	Алгоритм	Размерность признакового пространства			
		3	6	9	12
Pavia	гистограмма-дерево, $\nu_{min} = 35$	0.95	0.88	0.8	0.73
	алгоритм C4.5	0.86	0.92	0.94	0.95
Pavia University	гистограмма-дерево, $\nu_{min} = 35$	0.8	0.76	0.67	0.68
	алгоритм C4.5	0.7	0.75	0.77	0.78
Salinas	гистограмма-дерево, $\nu_{min} = 35$	0.88	0.85	0.81	0.79
	алгоритм C4.5	0.62	0.8	0.87	0.89
Botswana	гистограмма-дерево, $\nu_{min} = 35$	0.79	0.75	0.69	0.56
	алгоритм C4.5	0.54	0.79	0.82	0.81
Indian pines	гистограмма-дерево, $\nu_{min} = 35$	0.64	0.64	0.57	0.58
	алгоритм C4.5	0.49	0.57	0.61	0.64
Kennedy Space Center	гистограмма-дерево, $\nu_{min} = 35$	0.56	0.58	0.52	0.51
	алгоритм C4.5	0.55	0.69	0.73	0.81

Что касается алгоритма C4.5, то можно заметить, что с ростом количества компонент точность классификации возрастает. Особенностью алгоритма C4.5 является выбор наиболее информативного признака для построения разбиения на каждом шаге. Наиболее информативными являются признаки, обеспечивающие максимум энтропии. Такая стратегия позволяет не только сохранять достигнутое на малых размерностях качество классификации, но и увеличивать его с ростом размерности, учитывая дополнительные признаки только на наиболее глубоких уровнях иерархии. Таким образом, можно заключить, что алгоритм C4.5 целесообразнее применять в ситуации большого числа признаков, оказывающих существенно неравномерное влияние на результат. Также как и в случае с гистограммой-дерево алгоритм C4.5 хуже всего проклассифицировал изображение Kennedy Space Center и Indian pines. При этом сравнивая эти два изображения можно сделать вывод, что алгоритм C4.5 менее чувствителен к колебаниям дисперсии по каналам исходного изображения, чем к объёму обучающей выборки, поскольку для размерности 9 и выше вероятность верной классификации изображения Kennedy Space Center оказалась весьма высокой и составила 0,73 против 0,61 для изображения Indian pines.

Дальнейшие результаты демонстрируют сравнение объёмов данных и времени работы обоих алгоритмов в рассматриваемом диапазоне размерностей.

Средний объём памяти, требуемый для хранения дерева каждым из алгоритмов, приведен таблице 4.

Для алгоритма C4.5 в рассматриваемом случае расходы памяти на хранение структуры данных меньше. Однако, рост объемов памяти при сопоставимом росте размерности в 4 раза для предложенного алгоритма составляет от 1.8 до 4.5 раз, а для алгоритма C4.5 от 13.4 до 40.6 раз. Таким образом, дальнейший рост размерности может привести к тому, что использование алгоритма C4.5 может оказаться менее эффективным с точки зрения расходов памяти.

Оценка времени обучения производилась на 75% размеченных данных, использовавшихся в качестве обучающих выборок при оценке качества классификации. Определение времени классификации производилось на всем объёме данных изображения, то есть с использованием неразмеченных отсчётов. Тестирование производительности выполнялось на оборудовании со следующей конфигурацией: операционная система Windows 8, разрядность операционной системы 64-бит, объём оперативной памяти 8Гб, процессор Intel Core i5-3470. Предлагаемый алгоритм был реализован на языке C++ в среде Microsoft Visual Studio 2013. Алгоритм C4.5 рассматривался в виде реализации на языке C# в среде Microsoft Visual Studio 2013 в составе пакета Accord .Net Framework.

В таблице 5 приведено среднее время работы алгоритмов в секундах на этапе обучения и классификации. Данные таблицы 5 позволяют заключить, что алгоритм на основе гистограммы-дерева обеспечивает существенно более высокие скорости работы, чем алгоритм C4.5.

Таблица 4. Объем памяти в килобайтах, требуемый для хранения структуры данных классификатора

Алгоритм	Изображение	Размерность признакового пространства			
		3	6	9	12
гистограмма-дерево, $\nu_{min} = 35$	Pavia	392	1115	1537	1769
	Pavia University	168	334	451	370
	Salinas	307	520	689	677
	Kennedy Space Center	28	48	56	52
	Botswana	20	34	28	69
	Indian pines	73	125	136	131
алгоритм C4.5	Pavia	5	13.93	59.55	176.78
	Pavia University	5	12.68	50.24	182.48
	Salinas	5	14.09	53.76	118.51
	Kennedy Space Center	5	15.07	51.28	60.25
	Botswana	5	16.41	41.17	60.25
	Indian pines	5	16.08	61.63	134.43

Таблица 5. Среднее время обучения t_l, c и классификации t_c, c для исследуемых алгоритмов

Алгоритм	Изображение	Размерность признакового пространства							
		3		6		9		12	
		t_l, c	t_c, c	t_l, c	t_c, c	t_l, c	t_c, c	t_l, c	t_c, c
гистограмма-дерево с $\nu_{min} = 35$	Pavia	0.359	0.338	0.764	0.512	2.502	1.258	13.388	7.758
	Pavia University	0.129	0.109	0.214	0.152	0.521	0.323	2.831	1.973
	Salinas	0.15	0.10	0.24	0.13	0.47	0.22	1.758	0.684
	Kennedy Space Center	0.073	0.121	0.102	0.145	0.111	0.201	0.147	0.352
	Botswana	0.102	0.131	0.127	0.169	2.502	1.258	0.244	1.965
	Indian pines	0.059	0.046	0.077	0.06	0.521	0.323	0.263	0.091
алгоритм C4.5	Pavia	56.24	57.80	159.92	162.63	283.27	286.93	430.55	434.90
	Pavia University	16.04	16.64	47.31	48.24	90.17	91.37	152.16	153.66
	Salinas	30.26	30.54	94.38	94.88	170.84	171.51	268.88	269.71
	Kennedy Space Center	1.91	2.82	3.36	4.50	5.02	6.44	14.81	16.75
	Botswana	2.23	2.97	5.20	6.65	9.17	11.01	14.81	16.75
	Indian pines	8.49	8.54	19.59	19.67	34.63	34.74	60.28	60.42

Предложенный алгоритм относит отсчёты, не удовлетворяющие разбиению, к отсчётам с неопределённым классом объектов (белые точки на рисунках). Используемая же реализация C4.5 лишена данной возможности и классифицирует отсчёт хотя бы в один из классов, на которых алгоритм был обучен. Для визуализации результатов классификации на рис. 5 приведены примеры изображения Salinas, проклассифицированного при размерности вектора признаков равной трём.

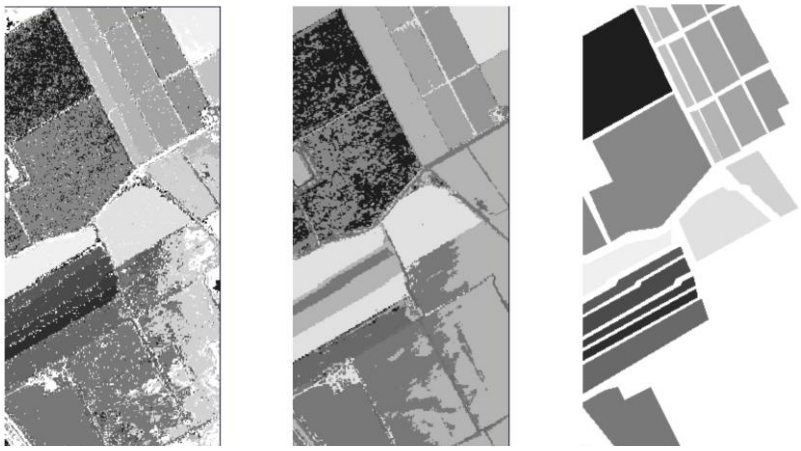

Рис. 5. Результаты классификации для трехмерных векторов признаков изображения Salinas, слева на право: гистограмма-дерево, алгоритм C4.5, маска классов (белые области соотетствют неопределённому номеру класса).

4. Выводы

В статье предложен и исследован алгоритм классификации на основе структуры данных гистограммы-дерева. Описанный алгоритм был сравнен с известным алгоритмом C4.5, реализующим построение классификатора в виде дерева решений. По сравнению с предлагаемым алгоритмом алгоритм C4.5 оказался менее эффективным по времени во всех случаях и более эффективным по памяти в большинстве случаев. Однако с увеличением размерности признакового пространства рост необходимых объёмов памяти для алгоритма C4.5 существенно выше, чем у предложенного алгоритма. Например, при возрастании размерности вектора признаков в 4 раза требуемый объем памяти для алгоритма C4.5 увеличился в 13,4 - 40,6 раз по сравнению с 1,8-4,5 раз для предложенного алгоритма. Данный факт позволяет сделать предположение о потенциальном превосходстве предложенного алгоритма с точки зрения затрачиваемых объемов памяти в случае большей размерности входных данных.

Проведённые эксперименты показали, что предложенный алгоритм имеет большую вероятность верной классификации, чем для алгоритм C4.5 при размерности вектора признаков 3 и 6, однако с ростом размерности вероятность верной классификации убывает. Данный факт обусловлен применением в качестве вектора признаков первых коэффициентов разложения методом главных компонент и малым объемом обучающей выборки. С увеличением размерности информативность добавляемых признаков снижается, что приводит к менее плотному расположению векторов признаков в пространстве большей размерности. Алгоритм C4.5, напротив, ориентирован на неравномерное влияние признаков. В ситуации равнозначности признаков предпочтительнее использовать алгоритм на основе гистограммы-дерева нежели алгоритм C4.5.

Благодарности

Работа выполнена при поддержке грантов РФФИ №16-29-09494 офи_м, № 16-37-00043 мол_а.

Литература

[1] Кузнецов, А. В. Сравнение алгоритмов управляемой поэлементной классификации гиперспектральных изображений/ А. В. Кузнецов , В.В.Мясников //Компьютерная оптика. – 2014. – Т. 38(3). – С. 494-502.

[2] Moradi, G. Using Statistical Histogram Based EM Algorithm for Apple Defect Detection / G. Moradi, M. Shamsi, M. H. Sedaaghi , S. Moradi // American Journal of Signal Processing . – 2012. – Vol. 2(2). – P. 10-14.

[3] Денисова, А. Ю. Алгоритмы построения гистограмм многоканальных изображений с использованием иерархических структур данных / А.Ю.Денисова, В.В.Сергеев // Компьютерная оптика. – 2016. – Т. 40(4). – С. 535-542.

[4] Schmidberger, G. Tree-based Density Estimation: Algorithms and Applications . – The University of Waikato. – 2009.

[5] Anderlini, L. Density Estimation Trees as fast non-parametric modelling tools/ L. Anderlini //Journal of Physics: Conference Series. – IOP Publishing, 2016. –Vol. 762(1). – P. 012042.

[6] Hyperspectral Remote Sensing Scenes [Electronic resource]. – Access mode: http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes (30.01.2017)

[7] Quinlan, J. R. C4. 5: programs for machine learning / J. R. Quinlan – Elsevier, 2014.

[8] Accord.Net Framework [Electronic resource] – Access mode: http://accord-framework.net/.(30.01.2017)