УДК 004.93'11 **DOI** 10.15622/sp.53.6

В.Г. Астафуров, А.В. Скороходов

ФОРМИРОВАНИЕ СИСТЕМЫ ИНФОРМАТИВНЫХ КЛАССИФИКАЦИОННЫХ ХАРАКТЕРИСТИК ПРИ РЕШЕНИИ ЗАДАЧИ КЛАССИФИКАЦИИ ОБЛАЧНОСТИ ПО СПУТНИКОВЫМ ДАННЫМ MODIS

Астафуров В.Г., Скороходов А.В. **Формирование системы информативных классификационных характеристик при решении задачи классификации облачности по спутниковым данным MODIS.**

Аннотация. Предложен алгоритм формирования системы эффективных классификационных характеристик, основанный на концепции усеченного перебора и использовании информации об индивидуальных показателях классификации при выборе гранул. Его вычислительная эффективность обеспечивается применением операций простого сравнения результатов классификации отдельных классов при выборе наиболее информативной гранулы на очередной итерации и использованием технологии параллельных вычислений на графических процессорах.

Рассмотрены известные методы усеченного перебора для формирования систем эффективных классификационных характеристик. Обсуждаются результаты поиска информативных признаков на примере решения задачи классификации облачности на основе применения вероятностной нейронной сети и информации о текстуре спутниковых снимков MODIS. Представлено описание используемого классификатора и статистического подхода к описанию текстуры изображений.

Определены наиболее эффективные классификационные характеристики облачности путем сравнения комбинаций текстурных признаков, полученных с помощью методов усеченного перебора. Показаны результаты исследования динамики изменения оценки правильно проклассифицированных образцов при выполнении различных алгоритмов поиска информативных признаков. Установлено, что разработанный в данной работе метод позволяет уменьшить разброс значений вероятности правильной классификации отдельных классов.

Ключевые слова: информативность, классификация, нейронная сеть, облачность, параллельные вычисления, текстурные признаки, усеченный перебор.

1. Введение. Эффективность решения задачи классификации зависит от успешного выполнения нескольких этапов. На первом из них формализуется классификационная модель и формируется репрезентативная обучающая выборка, состоящая из характерных образцов исследуемых объектов классификации. Второй этап заключается в выборе признаков для их описания и в формировании системы информативных классификационных характеристик. На третьем этапе осуществляется выбор типа классификатора и его настройка. Достоверность результатов классификации оценивается на основе тестовой выборки. Информативность используемых признаков описания объектов зависит не только от выбранного классификатора, но и от значений его параметров [1]. Поэтому целесообразно объединение и совместное выполнение второго и третьего этапов

решения задачи классификации. Таким образом, система эффективных классификационных характеристик выступает связующим элементом между экспертным представлением классификационной модели и ее математической интерпретацией, реализуемой с помощью методов интеллектуального анализа данных. При этом использование наиболее информативных признаков позволяет сократить время выполнения алгоритмов классификации и повысить их надежность.

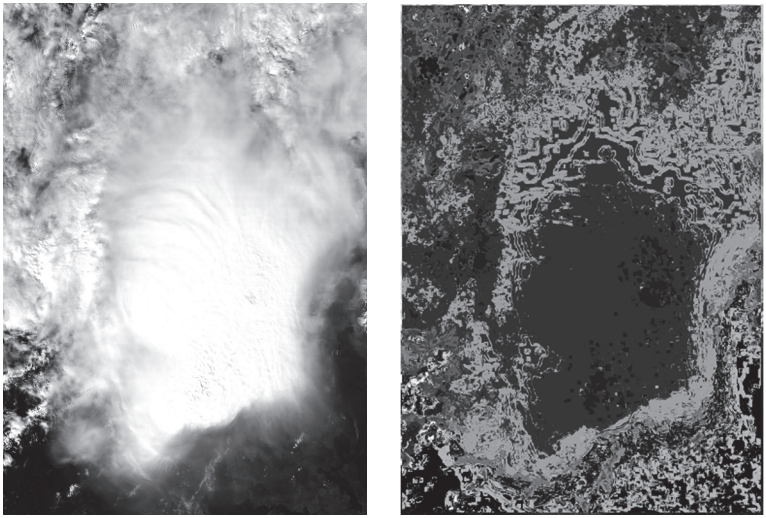
Система классификационных характеристик называется оптимальной, если ее применение позволяет достигнуть наилучших результатов классификации тестовой выборки на основе выбранного классификатора. При этом существует глобальное и локальное решение задачи поиска информативных признаков. Глобальная оптимальная система классификационных характеристик может быть сформирована путем полного перебора всех возможных комбинаций исследуемых признаков и оценки достоверности классификации. Однако такой подход наиболее трудоемкий и применим только к признаковым пространствам небольшой размерности [2]. Остальные алгоритмы формирования системы эффективных классификационных характеристик приводят к локальному решению, поскольку рассматривают ограниченное количество их комбинаций. Наилучшую производительность демонстрируют методы определения информативности признаков, основанные на использовании статистических характеристик, таких как корреляция между ними [3] или энтропия их комбинаций [4]. Главным недостатком такого подхода является то, что при формировании на его основе системы эффективных характеристик не учитывается достоверность результатов классификации. Поэтому, например, система некоррелированных признаков может содержать и неинформативные характеристики. Таким образом, этот подход является эффективным только лишь для сокращения размерности пространства признаков. Оптимальными по соотношению производительности и информативности системы классификационных характеристик могут быть сформированы на базе алгоритмов усеченного перебора, основанных на пошаговых процедурах улучшения результатов классификации.

К наиболее известным методам усеченного перебора относятся: Addition (Ad) [5], Deletion (Del) [6], комбинированный алгоритм AdDel [7] и его модификация GRAD [8]. Отличительной особенностью последних двух алгоритмов является то, что они позволяют одновременно сформировать системы эффективных классификационных характеристик и определить их оптимальное количество, при котором получаются наилучшие результаты

классификации. При этом переборные методы универсальны и могут применяться практически для любого алгоритма классификации [1]. Формирование системы информативных признаков с помощью методов усеченного перебора позволяет подобрать значения параметров классификаторов: количество нейронов (сети Кохонена), уровень соседства (алгоритм N-ближайших соседей), экспоненциальный вес (метод нечетких C-средних) и другие. К основным недостаткам методов усеченного перебора относятся: значительное снижение производительности при увеличении размерности пространства признаков и флуктуации результатов классификации отдельных классов при изменении параметров перебора.

В работе рассматриваются результаты разработки и исследования универсального высокопроизводительного алгоритма для формирования системы эффективных классификационных характеристик, основанного на методе усеченного перебора для решения задачи классификации облачности по спутниковым снимкам MODIS и его сравнение с существующими аналогами в данной области.

2. Исходные данные. В настоящее время задача классификации облачности по спутниковым данным по-прежнему является актуальной ввиду отсутствия ее решений в полном объеме. В существующих работах облака на изображениях из космоса надежно классифицируются только по 10-14 разновидностям [9-12], к которым относятся 10 основных форм облачности и несколько их подтипов. При этом в соответствии с метеорологическим стандартом [13] облака на наземных и судовых метеостанциях классифицируются по 27 разновидностям. В [1] нами показана возможность распознавания 25 сигнатур облачности с вероятностью правильной классификации (ВПК) 0,72 на основе использования вероятностной нейронной сети и информации о текстуре спутниковых снимков MODIS с пространственным разрешением 250 м в видимом диапазоне спектра (0,62-0,67 мкм), полученных в дневное время и при отсутствии снежного покрова. Однако алгоритм может быть адаптирован к данным и других спутниковых систем, учитывая их пространственное разрешение. Применение алгоритма Ad позволило не только сформировать систему эффективных классификационных характеристик, но и подобрать параметр сглаживания функции Гаусса. Однако при анализе результатов обработки тестовой выборки было установлено, что ВПК отдельных типов облаков изменяется от 0,16 до 1. При этом надежно выделяются только 11 разновидностей облачности с ВПК более 0,7. На рисунке 1 показан пример получаемых нами результатов классификации облаков по спутниковому снимку MODIS территории Греции от 14.09.2008 г. (UTC – 11:55).



а *б*

Рис. 1. Результаты классификации облачности по спутниковому снимку MODIS

Обучающая выборка сформирована с помощью методики сопоставления архивных данных сети наземных метеостанций со спутниковой съемкой MODIS и состоит из фрагментов характерных изображений 25 разновидностей облаков и двух типов подстилающей поверхности размером 21×21 пиксель [1]. При этом использовались метеонаблюдения, при которых фиксировалась только однослойная облачность или облака вертикального развития. Описание изображений облачности основано на статистических методах текстурного подхода, согласно которым анализируется пространственная взаимозависимость яркостей соседних пикселей: Gray-Level Co-occurrence Matrix (GLCM) [14], Gray-Level Difference Vector (GLDV) [15], Sum And Difference Histogram (SADH) [16] и One-Dimensional Signal Histogram (ODSH) [17]. Для уменьшения вычислительной сложности методов и снижения размерности матриц смежности (GLCM), векторов различия (GLDV), гистограмм сумм и разностей (SADH) применяется квантование яркостей на 20 уровней. При определении смежных пар пикселей учитываются угловые направления 0°, 45°, 90° и 135°. Поэтому полная система рассматриваемых в данной работе текстурных характеристик состоит из $N = 132$ признаков, которые приведены в таблице 1. Здесь группы из четырех одноименных признаков отличаются угловыми направлениями.

Таблица 1. Текстурные признаки для описания спутниковых снимков облачности

Обозначение	Текстурный признак	Обозначение	Текстурный признак
GLCM		$T_{73} – T_{76}$	Энтропия
$T_1 – T_4$	Второй угловой момент	$T_{77} – T_{80}$	Локальная однородность
$T_5 – T_8$	Энтропия	$T_{81} – T_{84}$	Контраст
$T_9 – T_{12}$	Максимальная вероятность	$T_{85} – T_{88}$	Кластерное затенение
$T_{13} – T_{16}$	Локальная однородность	$T_{89} – T_{92}$	Кластерная рельефность
$T_{17} – T_{20}$	Инверсия	*SADH*	
$T_{21} – T_{24}$	Дисперсия	$T_{93} – T_{96}$	Математическое ожидание
$T_{25} – T_{28}$	Контраст	$T_{97} – T_{100}$	Стандартное отклонение
$T_{29} – T_{32}$	Суммарное среднее	$T_{101} – T_{104}$	Второй угловой момент
$T_{33} – T_{36}$	Суммарная дисперсия	$T_{105} – T_{108}$	Контраст
$T_{37} – T_{40}$	Суммарная энтропия	$T_{109} – T_{112}$	Корреляция
$T_{41} – T_{44}$	Дифференциальная дисперсия	$T_{113} – T_{116}$	Энтропия
$T_{45} – T_{48}$	Дифференциальная энтропия	$T_{117} – T_{120}$	Локальная однородность
$T_{49} – T_{52}$	Корреляция	$T_{121} – T_{124}$	Кластерное затенение
$T_{53} – T_{56}$	Информационная мера -1	$T_{125} – T_{128}$	Кластерная рельефность
$T_{57} – T_{60}$	Информационная мера -2	*ODSH*	
GLDV		T_{129}	Первый начальный момент
$T_{61} – T_{64}$	Математическое ожидание	T_{130}	Энтропия
$T_{65} – T_{68}$	Стандартное отклонение	T_{131}	Энергия
$T_{69} – T_{72}$	Второй угловой момент	T_{132}	Вариация

В качестве классификатора применялась вероятностная нейронная сеть, преимуществами которой являются: возможность моделирования решающего правила практически любой сложности; фактическое отсутствие процедуры обучения; наличие единственного сглаживающего параметра функции активации σ и вероятностная интерпретация выходного отклика [18]. Основной недостаток такого типа сетей заключается в необходимости хранения

информации обо всей обучающей выборке в ее структуре, что обуславливает низкую производительность таких классификаторов. Однако простота и небольшое число используемых математических операций при их моделировании и применение технологии параллельных вычислений на графических процессорах NVIDIA CUDA нивелирует указанный недостаток [19].

Вероятностная нейронная сеть состоит из трех вычислительных слоев. Первый из них включает 5400 нейронов-образцов для каждой разновидности облачности и типа подстилающей поверхности в соответствии с объемом обучающей выборки. Откликами нейронов этого слоя являются значения уровней активности образцов. При этом для активации нейронов используется функция Гаусса. Второй слой сети содержит 27 нейронов-сумматоров по числу распознаваемых типов облачности и подстилающей поверхности. Отклики нейронов этого слоя представляют собой рассчитанные значения уровней активности каждого класса. Для их вычисления применяется функция Парзена, зависящая только от суммарных откликов нейронов предыдущего слоя для соответствующего класса. Выходной слой сети состоит из одного нейрона-консеквента, откликом которого является номер класса с наибольшим значением уровня активности. Архитектура вероятностной нейронной сети делает возможным ее реализацию в качестве функции-ядра при использовании технологии NVIDIA CUDA [20]. Использование методов параллельных вычислений существенно повышает эффективность алгоритмов усеченного перебора формирования систем информативных классификационных характеристик, описание которых представлено в следующем разделе данной работы.

3. Методы усеченного перебора. В основе всех методов усеченного перебора лежит итерационный подход, на каждом шаге которого ищется локальное оптимальное решение из ограниченной области. При этом общее количество проверяемых комбинаций признаков получается на несколько порядков меньше, чем при использовании метода полного перебора. Однако рассматриваемый подход может и не дать глобального оптимального решения, но полученный результат будет максимально приближен к нему. Таким образом, основной задачей при разработке методов усеченного перебора является повышение эффективности процедуры поиска локальных оптимальных решений путем выбора способа формирования комбинаций признаков и критериев их отбора. На данный момент известно несколько алгоритмов, реализующих данный подход: Ad [5], Del [6], AdDel [7] и GRAD [8]. Сравнение методов

усеченного перебора осуществлялось на тестовом наборе, состоящем из $N_T = 1350$ фрагментов различных разновидностей облачности и подстилающей поверхности, не включенных в обучающую выборку.

Алгоритм Ad базируется на последовательном переборе и добавлении наиболее информативных признаков к тестируемому набору. Основная суть методики заключается в следующем:

1) имеется исходная система исследуемых характеристик $T_S = \{T_1, T_2, T_3, ..., T_N\}$ и тестируемый набор $T_F = \varnothing$;

2) выполняется последовательный перебор признаков из T_S и их поочередное включение в T_F;

3) после каждого добавления признака производится классификация тестовой выборки с использованием системы признаков T_F и рассчитывается оценка ВПК:

$$E = \frac{N_R}{N_T},$$

где N_R — число правильно проклассифицированных образцов;

4) наиболее эффективный признак по результатам сравнения E остается в наборе T_F, а из T_S исключается. Таким образом, число тестируемых признаков $|T_F|$ увеличивается на единицу;

5) шаги 2-4 повторяются для оставшихся $N-1$ характеристик из T_S. При этом существуют несколько критериев остановки алгоритма. Один из них заключается в прекращении добавления новых признаков, если они не позволяют повысить ВПК тестовой выборки, достигнутую на предыдущем шаге. Второй заключается в искусственном ограничении максимального числа классификационных характеристик в T_F. Согласно третьему варианту алгоритм Ad выполняется пока $T_S \neq \varnothing$, что позволяет оценить, как меняется достоверность классификации тестовой выборки при использовании наборов с различными значениями T_F.

Аналогом Ad является алгоритм Del с той лишь разницей, что при последовательном переборе наименее информативный признак исключается из тестируемого набора согласно следующей последовательности действий:

1) на начальном этапе тестируемый набор включает все исследуемые признаки $T_F = \{T_1, T_2, T_3, ..., T_N\}$;

2) выполняется последовательный перебор характеристик из T_F и их поочередное удаление;

3) после каждого удаления признака проводится классификация тестовой выборки при использовании набора T_F и рассчитывается E;

4) наименее информативная характеристика, при которой E принимает минимальное значение, удаляется из T_F и далее в переборе не участвует;

5) шаги 2-4 повторяются для оставшихся $N-1$ признаков из T_F. Критерии остановки аналогичные, что и в методе Ad: удаление характеристик не улучшает результаты классификации тестовой выборки, максимальное число информативных признаков в итоговой системе задается заранее, $T_F \neq \varnothing$.

Методика AdDel является комбинированной и основана на попеременном исполнении алгоритмов Ad и Del, которое устраняет их зависимость от выбора начального информативного или неинформативного признака соответственно. Например, наиболее эффективная характеристика, выбранная на первой итерации метода Ad, может и не попасть в итоговую систему информативных признаков. Алгоритм AdDel представляется в следующем виде:

1) на начальном этапе тестируемый набор $T_F = \varnothing$ и исследуемый $T_S = \{T_1, T_2, T_3, ..., T_N\}$;

2) происходит выполнение алгоритма Ad и набираются N_A информативных признаков в T_F по результатам классификации тестовой выборки;

3) далее из полученного на шаге 2 тестируемого набора T_F методом Del исключается $N_D < N_A$ наименее эффективных характеристик;

4) шаги 2 и 3 повторяются для оставшихся $N - N_A + N_D$ признаков из T_S. Основные критерии остановки аналогичны таковым в методах Ad и Del. При этом наблюдения, отмеченные в [8], свидетельствуют о том, что при использовании алгоритма AdDel качество классификации вначале растет, а затем постепенно снижается за счет добавления малоинформативных признаков. Таким образом, можно определить оптимальное для данного метода число классификационных характеристик и использовать эту особенность в качестве критерия его остановки. В соответствии с практическими рекомендациями [8] число добавляемых N_A и удаляемых N_D

признаков инициализируется, как правило, в соотношении 2 к 1, что обеспечивает наибольшую эффективность метода AdDel.

Алгоритм GRAD является развитием AdDel, согласно которому в переборе участвуют уже не отдельные признаки, а некоторым образом сформированные их комбинации, называемые гранулами. Для их формирования прибегают к методу полного перебора, поскольку невозможно заранее оценить информативность того или иного сочетания исследуемых классификационных характеристик. При большом количестве исследуемых признаков следует ограничиваться двумя или тремя признаками. Основная же суть алгоритма GRAD заключается в следующем:

1) методом полного перебора формируются N_G гранул $T_G \subset T_S$, состоящих из заданного числа характеристик w;

2) имеется исходная система исследуемых признаков $T_S = \{T_1, T_2, T_3, ..., T_N\}$ и тестируемый набор $T_F = \varnothing$;

3) в тестируемый набор поочередно добавляются гранулы, сформированные на шаге 1. При этом обязательным условием является отсутствие любой характеристики из текущей гранулы в тестируемой подсистеме T_F;

4) после каждого добавления гранулы проводится классификация тестовой выборки при использовании набора T_F и оценивается ее ВПК;

5) признаки наиболее информативной гранулы остаются в подсистеме T_F, а из T_S исключаются;

6) шаги 3-5 повторяются в соответствии с заданным заранее числом добавляемых гранул N_G;

7) происходит включение стандартного алгоритма Del, в результате выполнения которого удаляется N_D наименее эффективных характеристик из T_F;

8) Шаги 3-7 повторяются для оставшихся $N - w \times N_C + N_D$ признаков. Критерии остановки идентичны таковым в методе AdDel. Вследствие возможного большого числа исследуемых классификационных характеристик N прибегают к сокращению количества используемых гранул. Для этого оценивают информативность каждой гранулы на этапе их формирования по результатам классификации тестовой выборки и в дальнейшем используют только наиболее эффективные из них (определенный процент от общего числа).

Соотношение же количества добавляемых и удаляемых характеристик является аналогичным, что и для метода AdDel.

В данной работе предлагается алгоритм являющийся развитием метода GRAD, основанный на учете индивидуальных показателей эффективности классификации отдельных классов, сокращенно GRAD-II (Individual Informativeness). Основная идея этого метода заключается в использовании гранул «способных» улучшить результаты классификации как можно большего числа рассматриваемых классов. Тем самым формируемая система эффективных классификационных характеристик, возможно, позволит не только повышать общую достоверность классификации, но и уменьшать разброс индивидуальных показателей распознавания рассматриваемых классов. Алгоритм GRAD-II имеет следующий вид:

1) методом полного перебора формируется N_G гранул $T_G^{(j)} \subset T_S$ при $j = 1, 2, ..., N_G$, состоящих из заданного числа характеристик w;

2) оцениваются общая ВПК $E\left(T_G^{(j)}\right)$ и результаты классификации каждого из $i = 1, 2, ..., K$ рассматриваемых классов $E_i\left(T_G^{(j)}\right)$ на основе тестовой выборки путем использования каждой гранулы $T_G^{(j)}$, сформированной на шаге 1;

3) при инициализации процедуры перебора в тестируемый набор T_F добавляется наиболее информативная гранула $T_G^{(j)}$, а исследуемая система признаков $T_S = \{T_1, T_2, T_3, ..., T_N\}$ за исключением характеристик из T_F;

4) производится поиск гранулы «способной» улучшить результаты классификации как можно большего числа рассматриваемых классов путем сравнения индивидуальных значений ВПК $E_i\left(T_F\right)$ и $E_i\left(T_G^{(j)}\right)$ при $j = 1, 2, ..., N_G - 1$:

$$H_j = \sum_{i=1}^{K} \begin{cases} 1, \text{ если } E_i\left(T_F\right) < E_i\left(T_G^{(j)}\right) \\ 0, \text{ в остальных случаях} \end{cases}$$

При этом рассматриваются только те комбинации признаков, для которых $T_G^{(j)} \subset T_S$. Таким образом, гранулы, в которых хотя бы один из признаков входит в T_F, из процедуры поиска исключаются;

5) комбинация с наибольшим значением H_j добавляется в тестируемый набор T_F, а соответствующие ей признаки удаляются из T_S;

6) проводится оценка результатов классификации при использовании T_F как общих $E(T_F)$, так и индивидуальных $E_i(T_F)$ оценок;

7) шаги 4-6 повторяются в соответствии с заданным заранее числом добавляемых гранул N_G;

8) происходит включение стандартного алгоритма Del, в результате выполнения которого удаляется N_D наименее эффективных характеристик из T_F;

9) шаги 4-8 повторяются для оставшихся $N - w \times N_C + N_D$ признаков. Критерии остановки и соотношение числа добавляемых и удаляемых характеристик идентичны таковым в методе GRAD. Вычислительная эффективность алгоритма GRAD-II достигается за счет выбора наилучшей гранулы путем простого сравнения индивидуальных показателей качества, проводимого на шаге 4, без использования процедуры классификации. При этом предварительная обработка, связанная с формированием гранул и расчетом значений ВПК отдельных классов, занимает практически одинаковое время, что и в методе GRAD.

4. Сравнительный анализ результатов поиска информативных признаков. Сравнение алгоритмов поиска эффективных классификационных характеристик, изложенных в предыдущем разделе, проводилось на примере решения задачи классификации облачности по спутниковым снимкам MODIS с использованием вероятностной нейронной сети и информации о текстуре изображений. Информативность сформированных комбинаций признаков оценивалась по результатам классификации тестовой выборки, состоящей из 1350 фрагментов спутниковых снимков 25 разновидностей облаков. При этом варьировался не только параметр сглаживания функции Гаусса σ, но и число добавляемых N_A и удаляемых N_D в методах AdDel, GRAD и GRAD-II характеристик.

На рисунке 2 показано влияние параметра σ на результаты классификации облачности при использовании метода Ad для формирования системы эффективных классификационных характеристик. Из рисунка 1 видно, что наиболее информативная комбинация признаков получена при $\sigma = 0,02$, что соответствует значению оценки ВПК $E = 0,713$. По мере увеличения параметра σ ВПК уменьшается и стабилизируется на уровне $E = 0,55$. Поэтому при анализе остальных методов поиска эффективных

классификационных характеристик значения $\sigma > 0{,}1$ не использовались. На рисунке 3 показаны результаты классификации тестовой выборки для различных значений тестируемых признаков $|T_F|$. Здесь и далее m обозначает номер итерации используемого метода усеченного перебора. Максимальное значение оценки ВПК достигается при $|T_F| = 30$. Следует отметить, что наибольший рост E наблюдается в интервале значений $|T_F|$ от 1 до 10. Добавленные при этом текстурные признаки обладают высокой информативностью. Далее в интервале от 10 до 30 качество классификации повышается незначительно, что свидетельствует о добавлении в систему малоинформативных признаков. При $|T_F| > 30$ к тестируемому набору присоединяются «шумовые» характеристики текстуры, о чем свидетельствует уменьшение оценки E для тестовой выборки.

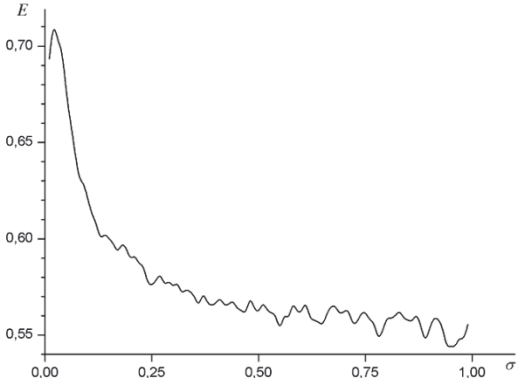


Рис. 2. Зависимость результатов классификации от параметра сглаживания

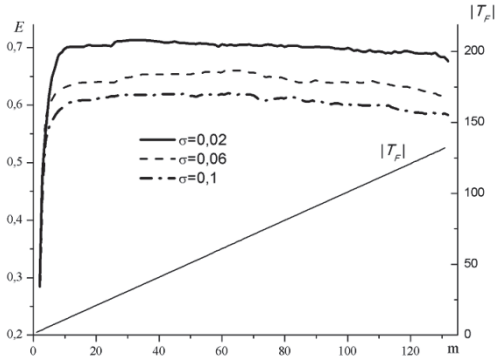


Рис. 3. Зависимость результатов классификации и числа тестируемых текстурных признаков для метода Ad

На рисунке 4 показаны результаты классификации тестовой выборки при различных значениях $|T_F|$ для алгоритма Del. Наибольшая ВПК $E = 0,705$ достигается при $|T_F| = 22$. Удаление «шумовых» признаков приводит к медленному росту E. Дальнейшее исключение малоинформативных характеристик влечет постепенное ухудшение оценки ВПК при $10 < |T_F| < 22$. Наиболее эффективные текстурные признаки удаляются, начиная с $|T_F| = 10$. Таким образом, алгоритмами Ad и Del достигнуты схожие результаты классификации тестовой выборки как по количеству информативных признаков, так и по динамике изменения оценки ВПК E.

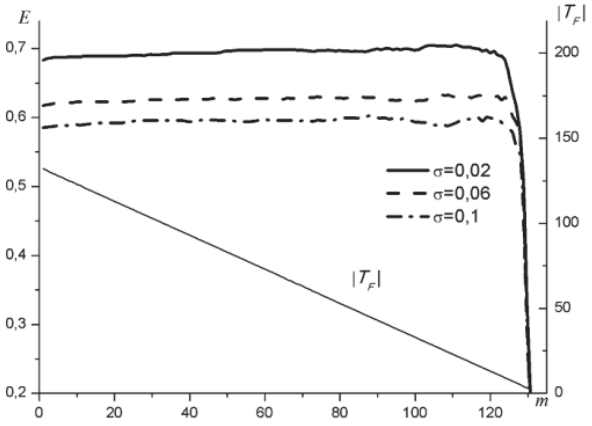


Рис. 4. Зависимость результатов классификации и числа тестируемых текстурных признаков для метода Del

Формирование систем информативных текстурных признаков путем применения метода AdDel проводилось при различном числе добавляемых N_A и удаляемых N_D признаков в соотношении 2 к 1. Наилучшие результаты классификации тестовой выборки, которые показаны на рисунке 5, были достигнуты при $N_A = 10$ и $N_D = 5$. При этом оценка ВПК $E = 0,716$ при $|T_F| = 25$. В общем случае изменение параметров N_A и N_D кардинальным образом не влияет на общую динамику оценки E, стабилизация которой происходит уже при

$|T_F| > 10$. По аналогии с методом Ad дальнейшее увеличение числа тестируемых признаков в T_F приводит вначале к медленному росту E, а затем к постепенному уменьшению этого показателя.

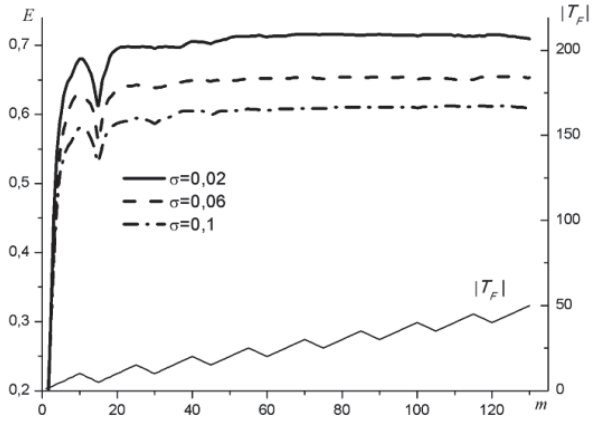


Рис. 5. Зависимость результатов классификации и числа тестируемых текстурных признаков в тестируемом наборе при использовании метода AdDel

Анализ результатов классификации тестовой выборки при использовании метода GRAD проводился путем варьирования не только числа удаляемых признаков N_D и добавляемых гранул N_C, но и их размеров w. Поскольку число исследуемых в работе текстурных признаков $N = 132$ достаточно велико, то рассматривались только двумерные ($w = 2$) и трехмерные ($w = 3$) комбинации классификационных характеристик. На рисунке 6 показаны наилучшие результаты обработки тестовой выборки, полученные при $N_C = 1$, $N_D = 1$ и $w = 2$. При этом была достигнута оценка ВПК $E = 0,721$ при $|T_F| = 25$. Как видно из рисунка 6, процесс увеличения E стабилизируется начиная с $|T_F| = 14$, что незначительно превосходит аналогичный показатель в методах Ad, Del и AdDel, что, видимо, связано с тем, что в T_F добавляются не отдельные признаки, а их комбинации. В остальном динамика изменения E типична в сравнении с рассмотренными алгоритмами.

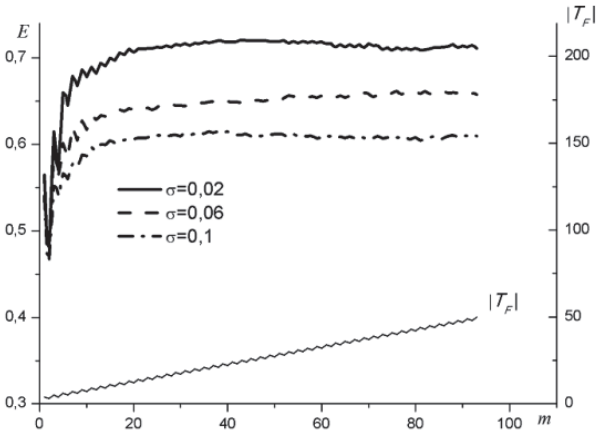


Рис. 6. Зависимость результатов классификации и числа тестируемых текстурных признаков для GRAD

Оценка результатов работы предложенного алгоритма формирования системы эффективных классификационных характеристик GRAD-II осуществлялась по схожей методике, что и при использовании GRAD: варьировалось число удаляемых признаков N_D и добавляемых гранул N_C, а также их размер w. Кроме того, проверена гипотеза о применении только наиболее информативных гранул при реализации алгоритма. На рисунке 7 показаны наилучшие результаты классификации тестовой выборки при использовании метода GRAD-II, достигнутые при $N_C = 3$, $N_D = 3$ и $w = 2$. При этом оценка ВПК составила $E = 0,724$ при $|T_F| = 44$. Рассмотрение только наиболее информативных гранул ухудшает результаты классификации до $E = 0,718$ при $|T_F| = 37$, $N_C = 4$, $N_D = 8$ и $w = 3$. Из рисунка 7 видно, что динамика изменений оценки ВПК E отличается от рассмотренных выше алгоритмов. При увеличении числа информативных признаков в T_F с 2 до 10 наблюдается значительный рост E. Дальнейшее добавление гранул в тестируемый набор постепенно повышает качество классификации, доходя до максимального значения E при значительном большем числе тестируемых признаков $|T_F|$, чем в других алгоритмах. При этом наблюдается сильная неравномерность изменения оценки ВПК, что предположительно связано с тем, что гранулы добавляются на

основании индивидуальных показателей достоверности $E_i\left(T_G^{(j)}\right)$, а не общей ВПК E.

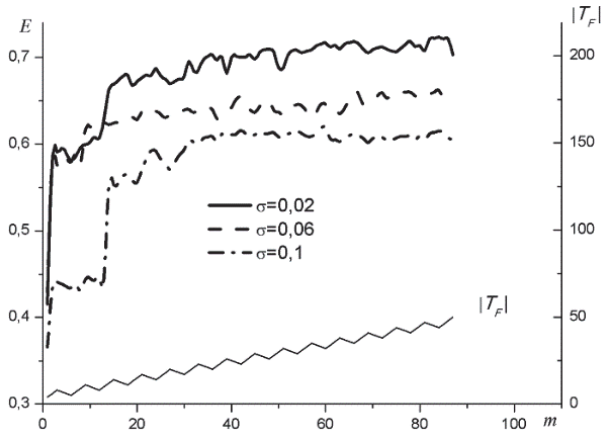


Рис. 6. Зависимость результатов классификации и числа тестируемых текстурных признаков для метода GRAD-II

Результаты применения методов усеченного перебора обобщены в таблице 2, из которой видно, что наиболее информативная система классификационных характеристик получена с помощью алгоритма GRAD-II. Оценка ВПК для тестовой выборки, достигнутая при использовании наихудшего набора, сформированного на основе метода Del, меньше на 0,02, что свидетельствует о схожей эффективности рассмотренных методов усеченного перебора. Можно предположить, что и оптимальная система классификационных характеристик, полученная путем полного перебора, позволит получить сравнимое значение оценки ВПК. Следует отметить, что наименьшее время формирования наборов информативных признаков t при использовании графического процессора GeForce GTX 780 достигается также при использовании алгоритма GRAD-II, поскольку выбор гранул происходит путем простого сравнения оценок ВПК отдельных типов облачности $E_i\left(T_F\right)$ и $E_i\left(T_G^{(j)}\right)$. При этом процедура формирования самих гранул занимает практически одинаковое время в методах GRAD и GRAD-II. Число классификационных характеристик, подобранных методов GRAD-II по сравнению с другими алгоритмами усеченного перебора максимально. Из рисунков 3-7

видно, что при значениях $|T_F| > 10$, ВПК тестовой выборки увеличивается менее чем на 0,05.

Таблица 2. Результаты формирования наборов информативных признаков

	Ad	*Del*	*AdDel*	*GRAD*	*GRAD-II*
E	0,713	0,705	0,716	0,721	0,724
$\|T_F\|$	30	22	25	25	44
t, с	$7{,}2\times10^3$	$7{,}2\times10^3$	$7{,}5\times10^3$	$15{,}2\times10^3$	$2{,}9\times10^3$
σ	0,02				0,03

Полученная на основе алгоритма GRAD-II система информативных текстурных признаков позволила уменьшить разброс значений ВПК отдельных типов облачности от $\min\limits_i E_i(T_F) = 0{,}16$ и $\max\limits_i E_i(T_F) = 1$ (метод Ad) до $\min\limits_i E_i(T_F) = 0{,}37$ и $\max\limits_i E_i(T_F) = 0{,}85$, что свидетельствует об эффективности предложенного подхода, учитывающего индивидуальные показатели классификации. Вопрос о том, какой из наборов информативных признаков является предпочтительным для решения задач многоклассовой классификации [21], с малым или большим разбросом значений вероятности правильного распознавания отдельных классов, остается открытым и требует дальнейшего рассмотрения.

Сравнивая наборы эффективных классификационных характеристик, сформированные путем применения рассматриваемых в работе методов перебора, можно выделить наиболее информативные текстурные признаки, которые входят во все эти наборы:

– информационная мера-1, вычисляемая для углового направления 90° по методу GLCM;
– математическое ожидание 90° (GLDV);
– стандартное отклонение 135° (GLDV);
– второй угловой момент 90° (GLDV);
– кластерное затенение 45° (GLDV);
– первый начальный момент (ODSH).

Следует отметить, что $2/3$ указанных текстурных признаков добавлялись в тестируемую комбинацию T_F одними из первых, обеспечивая наибольший прирост значений E. Наименее информативными классификационными характеристиками, которые не вошли ни в одну из сформированных систем, является большинство текстурных признаков, рассчитываемых по методу SADH.

5. Заключение. Предложен универсальный алгоритм формирования системы эффективных классификационных

характеристик GRAD-II, основанный на концепции усеченного перебора и использовании информации об индивидуальных показателях классификации отдельных классов $E_i\left(T_G^{(j)}\right)$ при выборе гранул. Вычислительная эффективность разработанного метода обеспечивается применением операций простого сравнения $E_i\left(T_G^{(j)}\right)$ и $E_i\left(T_F\right)$ при выборе наиболее информативной гранулы на очередной итерации без выполнения процедуры классификации.

Проведен сравнительный анализ алгоритмов формирования наборов эффективных классификационных характеристик применительно к задаче классификации облачности на основе использования вероятностной нейронной сети и информации о текстуре облаков на спутниковых снимках MODIS. Установлено, что применение рассмотренных в работе методов поиска информативных признаков позволило достигнуть схожих результатов классификации тестовой выборки (таблица 2). Можно предположить, что использование алгоритма полного перебора позволит незначительно улучшить достигнутые нами результаты классификации. На основании рисунков 3-7 выявлены три стадии изменения оценки ВПК: наибольшего прироста ($\left|T_F\right| < 10$), стабилизации ($10 < \left|T_F\right| < 30$) и последующего уменьшения E ($\left|T_F\right| > 30$). Также установлено, что применение разработанного алгоритма GRAD-II позволило уменьшить разброс значений ВПК отдельных типов облачности от $\min\limits_i E_i(T_F) = 0{,}16$ и $\max\limits_i E_i(T_F) = 1$ (метод Ad) до $\min\limits_i E_i(T_F) = 0{,}37$ и $\max\limits_i E_i(T_F) = 0{,}85$.

Все наборы эффективных классификационных характеристик, полученные различными методами усеченного перебора, содержат 6 общих наиболее информативных текстурных признаков: информационная мера-1 135° (GLCM); математическое ожидание 90° (GLDV); стандартное отклонение 135° (GLDV); второй угловой момент 90° (GLDV); кластерное затенение 45° (GLDV); первый начальный момент (ODSH). При этом $2/3$ из них добавлялись в тестируемую систему T_F на стадии наибольшего прироста E. Поэтому эти характеристики текстуры формируют «информативное ядро», которое вносит наибольший вклад в результаты классификации тестовой выборки. При этом большинство признаков, рассчитываемых методу SADH, следует отнести к неинформативным, поскольку они не вошли ни в одну из систем эффективных классификационных характеристик.

Алгоритмы на основе методов усеченного перебора представляют собой эффективный инструмент для формирования наборов информативных признаков близких к оптимальному, являются универсальными и могут использоваться совместно практически с любым классификатором. Вычислительная эффективность рассмотренных методов обеспечивается пошаговым улучшением результатов классификации за счет выбора наиболее информативных признаков и может быть повышена путем использования технологии параллельных вычислений NVIDIA CUDA.

Литература

1. *Астафуров В.Г., Курьянович К.В., Скороходов А.В.* Методы автоматической классификации облачности по спутниковым снимкам MODIS // Исследование Земли из космоса. 2016. № 4. С. 35–45.
2. *Загоруйко Н.Г.* Методы распознавания и их применение // М.: Изд-во «Советское радио». 1972. 208 с.
3. *Астафуров В.Г., Скороходов А.В.* Сегментация спутниковых снимков облачности по текстурным признакам на основе нейросетевых технологий // Исследование Земли из космоса. 2011. № 6. С. 10−20.
4. *Bankert R.L.* Cloud classification of AVHRR imagery in maritime regions using a probabilistic neural network // J. Appl. Meteor. 1994. vol. 33. pp. 909–918.
5. *Барабаш Ю.Л., Варский Б.В., Зиновьев В.Т.* Автоматическое распознавание образов // Киев: Изд-во КВАИУ. 1963. 173 с.
6. *Merill T., Green O.M.* On the effectiveness of receptions in recognition systems // IEEE Trans. Inform. Theory. 1963. vol. IT-9. pp. 11–17.
7. *Кутин Г.И.* Методы ранжировки комплексов признаков. Обзор // Зарубежная радиоэлектроника. 1981. № 9. С. 54–70.
8. *Загоруйко Н.Г.* Когнитивный анализ данных // Новосибирск: Академическое издательство ГЕО. 2013. 186 с.
9. *Jin W., Gong F., Zeng X., Fu R.* Classification of clouds in satellite imagery using adaptive fuzzy sparse representation // Sensors. 2016. vol. 16. no. 12. pp. 2153.
10. *Hiroshi S., Takahito I., Kouki M.* High-resolution cloud analysis information derived from Himawari-8 data // Meteorological sattelite center technical note. 2016. vol. 61. pp. 43–51.
11. *Tapakis R., Charalambides A.G.* Equipment and methodologies for cloud detection and classification: A review // Solar Energy. 2013. vol. 95. pp. 392–430.
12. *Волкова Е.В.* Оценки параметров облачного покрова, осадков и опасных явлений погоды по данным радиометра AVHRR с МИСЗ серии NOAA круглосуточно в автоматическом режиме // Современные проблемы дистанционного зондирования Земли из космоса. 2013. Т. 10. № 3. С. 66–74.
13. Федеральная служба по гидрометеорологии и мониторингу окружающей среды (Росгидромет). Код для оперативной передачи данных приземных метеорологических наблюдений с сети станций Росгидромета // М.: «Триада. лтд». 2013. 79 с.
14. *Haralick R.M., Shanmugam K, Dinstein I.* Textural features for image classification // IEEE Transactions on Systems, Man and Cybernetics. 1973. vol. SMC–3. no. 6. pp. 610–621.
15. *Weszka J.S., Dyer C.R., Rosenfeld A.* A comparative study of texture measures for terrain classification // IEEE Transaction on Systems, Man and Cybernetics. 1976. vol. SMC–6. no. 4. pp. 269–285.

16. *Unser M.* Sum and difference histograms for texture classification // IEEE Transaction on Systems, Pattern Analysis and Machine Intelligence. 1986. vol. PAMI–8. no. 1. pp. 118–125.
17. *Колодникова Н.В.* Обзор текстурных признаков для задач распознавания образов // Доклады Томского государственного университета систем управления и радиоэлектроники. 2004. Т. 9. № 1. С. 113–124.
18. *Specht D.F.* Probabilistic neural networks // Neural Networks. 1990. vol. 3. pp. 109–118.
19. *Savchenko A.V.* Pattern recognition and increasing of the computational efficiency of a parallel realization of the probabilistic neural network with homogeneity testing // Optical Memory and Neural Networks (Information Optics). 2013. vol. 22. no. 2. pp. 184–192.
20. *Скороходов А.В., Аксёнов С.В., Аксёнов А.В., Лайком Д.Н.* Использование различных вычислительных систем для решения задачи автоматической классификации облачности по спутниковым данным MODIS на основе вероятностной нейронной сети // Известия Томского политехнического университета. Инжиниринг георесурсов. 2016. Т. 327. № 1. С. 30–38.
21. *Student S., Pieter J., Fujarewicz K.* Multiclass classification problem of large-scale biomedical meta-data // Procedia Technology. 2016. vol. 11. pp. 938–945.

Астафуров Владимир Глебович — д-р физ.-мат. наук, профессор кафедры автоматизированных систем управления, Томский государственный университет систем управления и радиоэлектроники (ТУСУР), старший научный сотрудник группы атмосферной акустики, Федеральное государственное бюджетное учреждение науки Институт оптики атмосферы им. В.Е. Зуева Сибирского отделения Российской академии наук (ИОА СО РАН). Область научных интересов: разработка статистических методов анализа данных при решении задач классификации и распознавания образов. Число научных публикаций — 125. astafurov@iao.ru; пл. Академика Зуева, 1, Томск, 634055; р.т.: +7(3822)49-22-56, Факс: +7(3822)49-20-86.

Скороходов Алексей Викторович — к-т техн. наук, научный сотрудник группы атмосферной акустики, Федеральное государственное бюджетное учреждение науки Институт оптики атмосферы им. В.Е. Зуева Сибирского отделения Российской академии наук (ИОА СО РАН). Область научных интересов: разработка алгоритмов интеллектуальной обработки данных дистанционного зондирования Земли из космоса с использованием технологии искусственных нейронных сетей, методов нечеткой логики и кластерного анализа. Число научных публикаций — 45. vazime@yandex.ru; пл. Академика Зуева, 1, Томск, 634055; р.т.: +7(3822)49-22-56, Факс: +7(3822)49-20-86.

Поддержка исследований. Работа выполнена при финансовой поддержке РФФИ (проект № 16-37-60019 мол_а_дк).

DOI 10.15622/sp.53.6

V.G. ASTAFUROV, A.V. SKOROKHODOV

FORMATION OF A SET OF INFORMATIVE CLASSIFICATION FEATURES FOR SOLVING CLOUD CLASSIFICATION PROBLEM USING MODIS SATELLITE DATA

Astafurov V.G., Skorokhodov A.V. **Formation of a Set of Informative Classification Features for Solving Cloud Classification Problem using MODIS Satellite Data.**

Abstract. An algorithm for the formation of a set of effective classification features, based on the truncated search concept and the use of the information about individual classification indicators in the granules selection, is proposed. Its computational efficiency is ensured by the use of simple comparison operations of classification results of individual classes when choosing the most informative granule at the next iteration and using the parallel computing technology on graphics processing units.

Known methods of the truncated selection for the formation of sets of effective classification features are considered. The results of the informative features search are discussed through the example of solving the cloud classification problem on the basis of the application of a probabilistic neural network and the texture information of MODIS satellite imagery. A description of the used classifier and the statistical approach to describing the texture of images is given.

The most effective cloud classification characteristics are determined by comparing the combinations of textural features obtained by truncated selection methods. The study results of the change dynamics in the correctly classified clouds estimation when performing various algorithms for informative features searching are shown. It is established that the method, developed in this paper, makes it possible to reduce the variance of probability values of the correct classification of individual classes.

Keywords: informative, classification, neural network, cloud, parallel computing, texture features, truncated select methods.

Astafurov Vladimir Glebovich — Ph.D., Dr. Sci., professor of automation and data processing department, Tomsk State University of Control Systems and Radioelectronics (TUSUR), senior researcher of atmospheric acoustic division, V.E. Zuev Institute of Atmospheric Optics SB RAS (IAO SB RAS). Research interests: statistical methods of data analysis in solution of classification and pattern recognition problems. The number of publications — 125. astafurov@iao.ru; 1, Academician Zuev square, Tomsk, 634055, Russia; office phone: +7(3822)49-22-56, Fax: +7(3822)49-20-86.

Skorokhodov Aleksey Victorovich — Ph.D., researcher of atmospheric acoustic division, V.E. Zuev Institute of Atmospheric Optics SB RAS (IAO SB RAS). Research interests: development of algorithms for intellectual processing of remote sensing data from Earth observation using artificial neural networks technology, fuzzy logic methods and cluster analysis. The number of publications — 45. vazime@yandex.ru; 1, Academician Zuev square, Tomsk, 634055, Russia; office phone: +7(3822)49-22-56, Fax: +7(3822)49-20-86.

Acknowledgements. This research is supported by RFBR (grant 16-37-60019 mol_a_dk).

References

1. Astafurov V.G., Kuriyanovich K.V., Skorokhodov A.V. [Methods for automatic cloud classification from MODIS data]. *Issledovanie Zemli iz kosmosa – The study of the Earth from space.* 2016. no. 4. pp. 30–45. (In Russ.).
2. Zagorujko N.G. *Metody raspoznavanija i ih primenenie* [Detection Methods and Their Application]. Moscow: “Sovetskoe radio” Publ. 1972. 208 p. (In Russ.).

3. Astafurov V.G., Skorokhodov A.V. [Segmentation of satellite images by textural parameters based on neural network technologies]. *Issledovanie Zemli iz kosmosa – The study of the Earth from space.* 2011. vol. 6. pp. 10–20. (In Russ.).
4. Bankert R.L. Cloud classification of AVHRR imagery in maritime regions using a probabilistic neural network, *Journal of Applied Meteorology*. 1994. vol. 33. pp. 909–918.
5. Barabash Ju.L., Varskij B.V., Zinov'ev V.T. *Avtomaticheskoe raspoznavanie obrazov* [Automatic recognition of images]. Kiev: "KVAIU" Publ. 1963. 173 p. (In Russ).
6. Merill T., Green O.M. On the effectiveness of receptions in recognition systems. *IEEE Transactions Information Theory*. 1963. vol. IT-9. pp. 11–17.
7. Kutin G.I. [Methods ranking of signs complexes. Review]. *Zarubezhnaja radiojelektronika – International radioelectronics.* 1981. vol. 9. pp. 54–70. (In Russ.).
8. Zagorujko N.G. *Kognitivnyj analiz dannyh* [Cognitive data analysis]. Novosibirsk: "GEO" Academic Publ., 2013. 186 p. (In Russ.).
9. Jin W., Gong F., Zeng X., Fu R. Classification of clouds in satellite imagery using adaptive fuzzy sparse representation. *Sensors*. 2016. vol. 16. no. 12. pp. 2153.
10. Hiroshi S., Takahito I., Kouki M. High-resolution cloud analysis information derived from Himawari-8 data. *Meteorological sattelite center technical note.* 2016. vol. 61. pp. 43–51.
11. Tapakis R., Charalambides A.G. Equipment and methodologies for cloud detection and classification. A review. *Solar Energy*. 2013. vol. 95. pp. 392–430.
12. Volkova E.V. [Automatic estimation of cloud cover and precipitation parameters obtained by AVHRR NOAA for day and night conditions]. *Sovremennye problemy distancionnogo zondirovanija Zemli iz kosmosa – Modern problems of remote sensing of the Earth from space*. 2013. vol. 10. no. 3. pp. 66–74. (In Russ.).
13. Federal'naja sluzhba po gidrometeorologii i monitoringu okruzhajushhej sredy (Rosgidromet). *Kod dlja operativnoj peredachi dannyh prizemnyh meteorologicheskih nabljudenij s seti stancij Rosgidrometa* [Code for the rapid transmission of data from surface meteorological observation network of Roshydromet stations]. Moscow: "Triada ltd." Publ. 2013. 79 p. (In Russ.).
14. Haralick R.M., Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*. 1973. vol. SMC–3. no. 6. pp. 610–621.
15. Weszka J.S., Dyer C.R., Rosenfeld A. A comparative study of texture measures for terrain classification, *IEEE Transaction on Systems, Man and Cybernetics.* 1976. vol. SMC–6. no. 4. pp. 269–285.
16. Unser M. Sum and difference histograms for texture classification. *IEEE Transaction on Systems, Pattern Analysis and Machine Intelligence*. 1986. vol. PAMI–8. no. 1. pp. 118–125.
17. Kolodnikova N.V. [Review of textural features for pattern recognition problems]. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravlenija i radiojelektroniki – Reports of the Tomsk State University of Control Systems and Radio Electronics*. 2004. vol. 9. no. 1. pp. 113–124. (In Russ.).
18. Specht D.F. Probabilistic neural networks. *Neural Networks.* 1990. vol. 3. pp. 109–118.
19. Savchenko A.V. Pattern recognition and increasing of the computational efficiency of a parallel realization of the probabilistic neural network with homogeneity testing. *Optical Memory and Neural Networks (Information Optics)*. 2013. vol. 22. no. 2. pp. 184–192.
20. Skorokhodov A.V., Aksjonov S.V., Aksjonov A.V., Lajkom D.N. [Using different computing systems to solve the automatic cloud classification problem on MODIS satellite data by probabilistic neural network]. *Izvestija Tomskogo politehnicheskogo universiteta. Inzhiniring georesursov – Bulletin of the Tomsk Polytechnic University. Engineering georesources.* 2016. vol 327. no. 1, pp. 30–38. (In Russ.).
21. Student S., Pieter J., Fujarewicz K. Multiclass classification problem of large-scale biomedical meta-data. *Procedia Technology.* 2016. vol. 11. pp. 938–945.