Федеральное государственное бюджетное учреждение науки
Институт вычислительных технологий Сибирского отделения
Российской академии наук

На правах рукописи

Синявский Юрий Николаевич

НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ И ПРОГРАММНО-АЛГОРИТМИЧЕСКИЙ ИНСТРУМЕНТАРИЙ ДЛЯ СЕГМЕНТАЦИИ МУЛЬТИСПЕКТРАЛЬНЫХ СПУТНИКОВЫХ ИЗОБРАЖЕНИЙ

05.13.18 – математическое моделирование, численные методы
и комплексы программ

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата технических наук

Научный руководитель:
к.ф.-м.н., доцент Пестунов Игорь Алексеевич

Новосибирск – 2020

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В последние десятилетия в области создания и развития средств и технологий дистанционного зондирования Земли наблюдается стремительный прогресс. Пространственное и спектральное разрешение съёмочной аппаратуры повышается, точность орбитальной привязки снимков постоянно улучшается. Кроме того, с каждым годом растёт число запускаемых космических аппаратов и, как следствие, наблюдается лавинообразный рост получаемых объёмов данных. В дополнение к этому, упрощается процедура получения спутниковых данных рядовых потребителей.

Спутниковые изображения в настоящее время активно используются при изучении обширных и труднодоступных территорий (в задачах, связанных с сельским и лесным хозяйством, картированием, экологией и охраной окружающей среды, прогнозированием и ликвидацией последствий чрезвычайных ситуаций и др.), поскольку эти данные, зачастую, являются единственным источником оперативной и объективной информации [1-3]. Поэтому необходимо постоянно совершенствовать алгоритмический и программный инструментарий для их обработки и анализа.

Одним из важнейших этапов анализа цифровых изображений является сегментация [4]. Она заключается в разбиении изображения на сегменты на основе однородности (похожести) их спектральных и/или пространственных (текстура, форма, размер и др.) характеристик. Сегментация преследует две основных цели [5]: (1) декомпозиция изображения на части, удобные для дальнейшего анализа и (2) группировка пикселей в более высокоуровневые и информативные структуры.

При сегментации спутниковых изображений объектами изучения являются отдельные элементы разрешения (пиксели), характеризующие небольшие участки поверхности Земли. При съёмке в нескольких спектральных диапазонах каждому пикселю ставится в соответствие набор спектральных характеристик (вектор признаков), который можно рассматривать как точку многомерного евклидова пространства. Эти признаки по своей природе являются случайными. Действительно,

два участка земной поверхности с идентичными свойствами в момент съёмки могут отражать по-разному, поскольку коэффициент отражения сильно зависит от внешних факторов (увлажнённость и освещённость участков, состав атмосферы над ними и т.п.). Случайный характер признаков также обуславливается шумами, неизменно присутствующими в каналах связи при передаче данных, а также предварительной обработкой снимков. Поэтому при обработке спутниковых изображений задачу сегментации зачастую приходится решать при отсутствии каких-либо априорных сведений о количестве классов и их вероятностных характеристиках. Это значительно затрудняет применение параметрических методов (а при отсутствии информации о виде плотностей распределения вероятности их корректное применение невозможно вовсе). Для описания реальных структур данных, представленных на спутниковых изображениях, наиболее подходящими являются скользящие локально-параметрические модели, лежащие в основе непараметрических методов кластеризации [6]. Эффективность такого рода алгоритмов подтверждена многочисленными экспериментальными исследованиями. Однако они не находят широкого применения при решении задач, связанных с обработкой спутниковых изображений, ввиду неприемлемо высокой вычислительной трудоёмкости.

При решении практических задач зачастую используются распространённые пакеты программ, предназначенные для анализа спутниковой информации (ERDAS Imagine, ITTVIS ENVI, eCognition, ESRI ArcGIS, QGIS, SAGA GIS и др.). В их состав входят традиционные, но, зачастую, устаревшие методы кластеризации данных. Эти методы реализуют лишь некоторые из стандартных подходов к классификации и не учитывают специфику спутниковых изображений. В то же время, более эффективные модели данных остаются невостребованными.

Таким образом, наряду с созданием новых методов и алгоритмов для анализа и обработки спутниковых изображений, необходимо уделить внимание разработке удобного механизма обеспечения доступа к ним.

Обеспечение оперативного доступа пользователей к новым алгоритмам обработки спутниковых изображений – сложная и трудоёмкая задача. При реализации

алгоритмов в виде модулей для автономных программных пакетов возникают следующие трудности:

- закрытость и высокая стоимость коммерческих пакетов;
- ограниченная базовая функциональность большинства открытых пакетов обработки;
- необходимость написания модулей расширения на внутреннем языке программирования (зачастую, узкоспециализированном и/или слабо распространённом);
- необходимость участия пользователя в процессе внедрения алгоритма (особенно актуально при большом числе пользователей).

Каждый программный пакет предназначен для решения некоторого узкого круга задач, поэтому базовая функциональность (а также пользовательский интерфейс) разных пакетов существенно различается. Следовательно, для внедрения алгоритма, который можно применить при решении широкого спектра задач, потребуется разработка нескольких модулей расширения для разных пакетов (или даже нескольких модулей для разных версий одного пакета), что является очень трудоёмким процессом.

В сложившейся ситуации наиболее перспективным способом обеспечения доступа к алгоритмам обработки является построение распределённых геоинформационных систем [7], предоставляющих функции обработки в виде стандартизованных веб-сервисов, которые могут быть использованы как открытыми, так и коммерческими программными пакетами. Такого рода системы предоставляют пользователю, работающему в произвольном пакете программ, возможность подобрать набор алгоритмов для получения оптимального решения поставленной задачи.

Развитию распределённых геоинформационных систем способствуют такие глобальные процессы, как:

- создание систем поиска по пространственно распределённым хранилищам спутниковых данных (см., например, [8]);
- внедрение стандартов предоставления пространственных данных [9, 10] и алгоритмов для их обработки [11];

- объединение крупнейших разработчиков программного обеспечения в консорциумы по стандартизации (в области геоинформатики создан Консорциум открытых ГИС – Open Geospatial Consortium, OGC [12]);
- развитие технологий поиска данных по их стандартизованным описаниям (метаданным) [13];
- постепенный перенос результатов различных исследований, имеющих географическую привязку (баз данных, карт, результатов полевых наблюдений и др.), в цифровые хранилища данных и организация доступа к ним [14].

Факты, изложенные выше, свидетельствуют об актуальности задачи и наличии предпосылок для разработки алгоритмов анализа и обработки спутниковых изображений, а также удобного механизма для их внедрения.

Целью диссертационной работы является разработка эффективных непараметрических алгоритмов сегментации спутниковых изображений и современной платформы для стандартизованного доступа к ним.

Для достижения этой цели необходимо решить следующие основные задачи.

- провести анализ алгоритмов кластеризации данных применительно к задаче сегментации мультиспектральных спутниковых изображений;
- разработать, теоретически обосновать и программно реализовать вычислительно эффективные непараметрические алгоритмы кластеризации для сегментации спутниковых изображений;
- выполнить экспериментальное сравнение разработанных алгоритмов с описанными в литературе на модельных данных и мультиспектральных изображениях;
- создать современную платформу для обеспечения стандартизованного доступа к алгоритмам сегментации мультиспектральных изображений, удобную как разработчикам, так и потенциальным пользователям.

Объектом исследования являются данные, которые представлены в виде мультиспектрального изображения, полученного при помощи многозональной оптико-электронной спутниковой системы.

Методы исследования опираются на современные информационно-вычислительные технологии, предусматривающие использование адекватных математических моделей изучаемого объекта и эффективных вычислительных алгоритмов. Для анализа мультиспектральных изображений используются методы теории вероятности и математической статистики. Для исследования разработанных алгоритмов применяются методы Монте – Карло. Программная реализация алгоритмов выполняется в рамках парадигмы объектно-ориентированного программирования с использованием стандартных паттернов проектирования Interface, Bridge, Prototype, Iterator и Template Method.

Экспериментальные исследования выполняются с использованием пакета программ «Image Processing Toolkit», разработанного автором.

На защиту выносятся следующие положения, соответствующие пунктам 3 («разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий»), 4 («реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента») и 8 («разработка систем компьютерного и имитационного моделирования») паспорта специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ» (технические науки).

1. Вычислительно эффективный алгоритм кластеризации MeanSC для сегментации мультиспектральных изображений, разработанный на основе непараметрических оценок Розенблатта – Парзена с учётом характерных особенностей спутниковых снимков.
2. Непараметрический ансамблевый алгоритм кластеризации EMeanSC для сегментации мультиспектральных изображений, построенный на основе согласованной матрицы различий.
3. Пакет программ «Image Processing Toolkit», включающий оригинальный программно-алгоритмический инструментарий для сегментации мультиспектральных изображений.

4. Набор стандартизованных веб-сервисов на основе предложенных алгоритмов и программного обеспечения с открытым исходным кодом.

Научная новизна диссертационной работы состоит в следующем.

1. Предложен вычислительно эффективный непараметрический алгоритм кластеризации MeanSC на основе оценок плотности Розенблатта – Парзена для сегментации мультиспектральных спутниковых изображений. Эффективность достигается за счёт введения сеточной структуры в пространстве признаков и переходу к рабочей выборке значительно меньшего объёма, в которой гарантированно содержатся представители всех классов, присутствующих на изображении. Сеточная структура впервые использована для повышения вычислительной эффективности поэлементного алгоритма кластеризации.
2. Предложен подход к построению ансамбля непараметрических алгоритмов кластеризации, основанных на оценках плотности Розенблатта – Парзена, с помощью согласованной матрицы различий. В рамках этого подхода на основе непараметрического алгоритма MeanSC создан ансамблевый алгоритм кластеризации EMeanSC, позволяющий обеспечить простоту настройки параметров и обработку мультиспектральных спутниковых изображений в диалоговом режиме.
3. На основе предложенных алгоритмов кластеризации разработаны методы разделения формаций лесной растительности с близкими спектрально-яркостными характеристиками и обнаружения усыхающих древостоев по мультиспектральным изображениям. Эти методы позволяют обеспечить качественное выделение мелких и сильно пересекающихся классов, которые не обнаруживаются при использовании традиционных методов автоматизированный обработки.

Практическая значимость полученных результатов обусловлена следующим.

Разработанные алгоритмы превосходят описанные в литературе по качеству классификации и/или вычислительной эффективности, что позволяет повысить эффективность сегментации спутниковых изображений в условиях малой априор-

ной информации при решении задач, связанных с исследованием и оценкой состояния территорий по данным дистанционного зондирования Земли. Алгоритмы внедрены в геоинформационную систему с открытым исходным кодом GRASS GIS, а также оформлены в виде стандартизованных веб-сервисов, что позволяет обеспечить доступ к ним по протоколу WPS.

Основные результаты работы были использованы при выполнении междисциплинарных интеграционных проектов СО РАН №№ 3 (2003-2005 гг.), 86 (2006-2008 гг.), 50 (2009-2011 гг.), проектов РФФИ (№№ 09-07-12087-офи_м, 11-07-12083-офи_м, 11-07-00202, 11-07-00346, 14-07-31320-мол_а, 18-37-00492-мол_а), международного гранта фонда «Научный потенциал» («Human Capital Foundation», 2006 г.) № 66, а также гранта мэрии г. Новосибирска № 09-09 (2009-2010 гг.).

Разработанный программно-алгоритмический инструментарий передан в Институт почвоведения и агрохимии СО РАН, где используется при крупномасштабном картографическом моделировании структурной организации растительности и почвенного покрова, что подтверждено актом об использовании.

Достоверность результатов обеспечивается корректным применением используемых методов и подтверждается проведенными экспериментальными исследованиями на модельных и прикладных задачах.

Представление работы. Результаты работы обсуждались на следующих научных мероприятиях: Международная конференция «Вычислительные и информационные технологии в науке, технике и образовании» (Павлодар, Казахстан, 2006); Всероссийская конференция «Современные методы математического моделирования природных и антропогенных катастроф» (Барнаул, 2007); Всероссийская конференция «Обработка пространственных данных и дистанционный мониторинг природной среды и масштабных антропогенных процессов» (Барнаул, 2013); Всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям (Красноярск, 2006; Томск, 2013); Международная конференция «Automation, Control and Information Technology» (Новосибирск, 2005; Новосибирск, 2010); Всероссийская конференция «Математические

методы распознавания образов» (Суздаль, 2009); Международная выставка и научный конгресс «ГЕО-Сибирь» (Новосибирск, 2010; Новосибирск, 2018); Всероссийская конференция с международным участием «Обработка пространственных данных в задачах мониторинга природных и антропогенных процессов» (Бердск, 2019); Международная научная конференция «Региональные проблемы дистанционного зондирования Земли» (Красноярск, 2014; Красноярск, 2018); объединенный семинар «Информационно-вычислительные технологии» в ИВТ СО РАН (Новосибирск, 2005-2015).

Публикации. По теме диссертации опубликовано 29 печатных работ, в том числе 7 статей в изданиях из Перечня ВАК, 3 – в изданиях, индексируемых в WoS и Scopus, 8 – в других рецензируемых журналах, 4 – в трудах и 7 – в тезисах международных и всероссийских конференций; получено 5 свидетельств о государственной регистрации программ для ЭВМ.

Личный вклад автора. Автор принимал активное участие в разработке методов и алгоритмов, а также в интерпретации результатов. Алгоритмы сегментации разработаны автором совместно с Пестуновым И. А. и Бериковым В. Б. Проектирование и программная реализация разработанных алгоритмов, создание пакета программ «Image Processing Toolkit» и проведение численных экспериментов, а также реализация алгоритмов в виде модулей для GRASS GIS и в виде стандартизованных веб-сервисов, выполнены автором лично.

Структура и объём работы. Текст диссертации состоит из введения, пяти глав, заключения, списка цитируемой литературы из 154 наименований и трёх приложений. Полный объем диссертации составляет 151 страницу, включая 49 рисунков и 7 таблиц.

ГЛАВА 1. СОВРЕМЕННОЕ СОСТОЯНИЕ ПРОБЛЕМЫ

1.1. Содержательная постановка задачи кластеризации

Методы кластеризации данных активно используются во многих областях науки. В их разработке принимают участие математики, информатики, биологи, социологи, медики, психологи и другие специалисты, которым необходимо анализировать различные результаты наблюдений. Термин «кластеризация данных» («data clustering») впервые появился в заголовке статьи 1954 года, посвящённой обработке антропологических данных [15]. В различных областях знаний кластеризации соответствуют разные термины: автоматическая классификация, классификация без обучения (без учителя), классификация с самообучением, таксономия, группировка, стратификация, типизация и др.

Содержательная постановка задачи кластеризации заключается в следующем. Пусть имеется некоторая выборка объектов $\Omega = \{\omega^{(1)}, \dots, \omega^{(N)}\}$, сформированная в результате отбора представителей из некоторой генеральной совокупности. Каждый объект исходной выборки $\omega^{(i)}$ характеризуется вектором признаков $x^{(i)} = x\left(\omega^{(i)}\right) = (x_1^{(i)}, \dots, x_k^{(i)}) \in R^k$. Иногда, в силу особенностей решаемой задачи, выборка Ω описывается матрицей коэффициентов попарного сходства/различия. Задача кластеризации заключается в разбиении выборки на сравнительно небольшое, заранее известное или нет, число $M \geq 2$ групп объектов (кластеров) так, чтобы элементы одного кластера были как можно более схожи, а элементы из разных кластеров существенно различались по заданному критерию сходства/различия.

Кластеризации обычно предшествует решение двух задач.

1. *Выбор меры схожести (или различия) объектов.* Мера схожести выбирается исходя из особенностей решаемой задачи и используется для определения расстояния между объектами. От неё напрямую зависят форма, размер и другие структурные характеристики выделяемых кластеров. Если группировка происходит поэтапно, то для групп объектов тоже необходимо задать способ вычисления

расстояния, например по принципу «дальнего соседа» (позволяет выделять компактные кластеры сферической формы), «ближнего соседа» (удобен при выделении концентрически расположенных кластеров, а также кластеров вытянутой формы) и др.

2. *Выбор или выделение признаков.* Зачастую, при решении практических задач очень сложно с уверенностью определить, какие признаки действительно важны для выделения классов интереса, поэтому исследователи включают как можно больше потенциально информативных факторов. Это может привести к эффекту, называемому «проклятьем размерности» («the curse of dimensionality») [16], когда с ростом размерности разница между наиболее похожей парой объектов становится сравнима с разницей между самыми непохожими. Поэтому при наличии большого числа признаков целесообразно сначала выбрать наиболее эффективную для решения конкретной задачи подсистему переменных. Этот процесс называется выбором признаков (feature selection) [17]. В некоторых задачах исходные признаки не позволяют выделить кластеры требуемого качества. В этом случае целесообразно использовать выделение признаков (feature extraction) [18], при котором исходная система переменных преобразуется в более подходящую.

1.2. Уровни априорной информации

При разработке статистических методов обработки и анализа изображений под априорной информацией понимаются любые сведения об изображении, имеющиеся до выполнения обработки [19]. Они могут быть более или менее полными. В зависимости от количества и качества априорной информации, различают три уровня неопределённости (см., напр. [20]):

1) байесовская неопределённость: функция плотности распределения вероятности известна полностью (точно известны вид плотности распределения вероятности и численные значения всех фигурирующих в нём коэффициентов);
2) параметрическая неопределённость – функция плотности распределения определена с точностью до значений параметров;

3) непараметрическая неопределённость – параметризованная структура функции плотности распределения вероятности неизвестна, имеется информация только о некоторых качественных её свойствах (например, сведения о непрерывности или дифференцируемости).

В данной работе задача сегментации изображений решается в условиях малой априорной информации (и, как следствие, неизвестного числа классов) с учётом следующих характерных особенностей.

1.3. Особенности задачи сегментации мультиспектральных спутниковых изображений

Задача сегментации мультиспектральных спутниковых изображений обладает четырьмя характерными особенностями.

Первая особенность заключается в большом объёме обрабатываемых данных. Современные спутниковые изображения, как правило, содержат порядка 10^6–10^7 элементов разрешения (пикселей).

Вторая особенность выражается в ограниченности диапазонов изменения значений спектральных признаков, обусловленной фиксированным числом уровнем квантования выходного сигнала съёмочной аппаратуры.

Третья особенность – недостаток (а иногда и полное отсутствие) априорной информации о количестве и вероятностных характеристиках классов, присутствующих на изображении. При исследовании природных объектов получение априорной информации зачастую связано со значительными и не всегда оправданными затратами ресурсов. Особенно остро эта проблема ощущается при исследовании обширных и труднодоступных территорий, когда спутниковые изображения являются единственным источником актуальной и объективной информации.

Четвёртая особенность рассматриваемой задачи заключается в присутствии на изображениях «шума» и выбросов, обусловленных особенностями съёмочной аппаратуры, условиями съёмки, присутствием облаков и др.

Для учёта указанных особенностей алгоритмы кластеризации должны обеспечивать: 1) возможность выделять кластеры разной структуры (формы, размера,

плотности) в условиях малой априорной информации (в том числе неизвестного числа кластеров); 2) низкую вычислительную сложность, позволяющую обрабатывать спутниковые изображения в диалоговом режиме; 3) возможность обрабатывать данные, содержащие «шум» и выбросы; 4) простоту настройки параметров.

К настоящему времени разработано большое количество различных алгоритмов кластеризации и их модификаций. Опубликовано множество обзорных статей [21-38] и монографий, полностью [39-46] или частично [5, 47-50] посвящённых кластеризации. В данной главе исследована возможность применения существующих алгоритмов кластеризации для сегментации спутниковых изображений.

1.4. Обзор существующих алгоритмов кластеризации

По способу выделения кластеров все алгоритмы автоматической классификации можно разделить на две большие группы – иерархические и неиерархические. Иерархические (hierarchical) алгоритмы позволяют обнаружить вложенную структуру кластеров. Для этого они строят либо дерево кластеров, называемое дендрограммой, либо так называемую диаграмму достижимости [51] (по которой можно построить дендрограмму [52]). Неиерархические алгоритмы вычисляют кластеры исходя из оптимизации некоторого заранее заданного (явно или неявно) критерия качества.

Неиерархические алгоритмы можно условно разделить на три большие группы: методы разбиений, плотностные методы и сеточные методы. Кроме того, можно выделить обособленную группу неиерархических алгоритмов, называемую нейронными сетями [53].

Перечисленные группы с примерами алгоритмов представлены на рисунке 1.1. Стоит заметить, что это разбиение условно; часто алгоритмы разрабатываются в рамках комбинации нескольких подходов. На рисунке 1.1 такие алгоритмы (например, DENCLUE или CLIQUE) отнесены одновременно к нескольким группам.

Рассмотрим перечисленные группы алгоритмов более подробно.

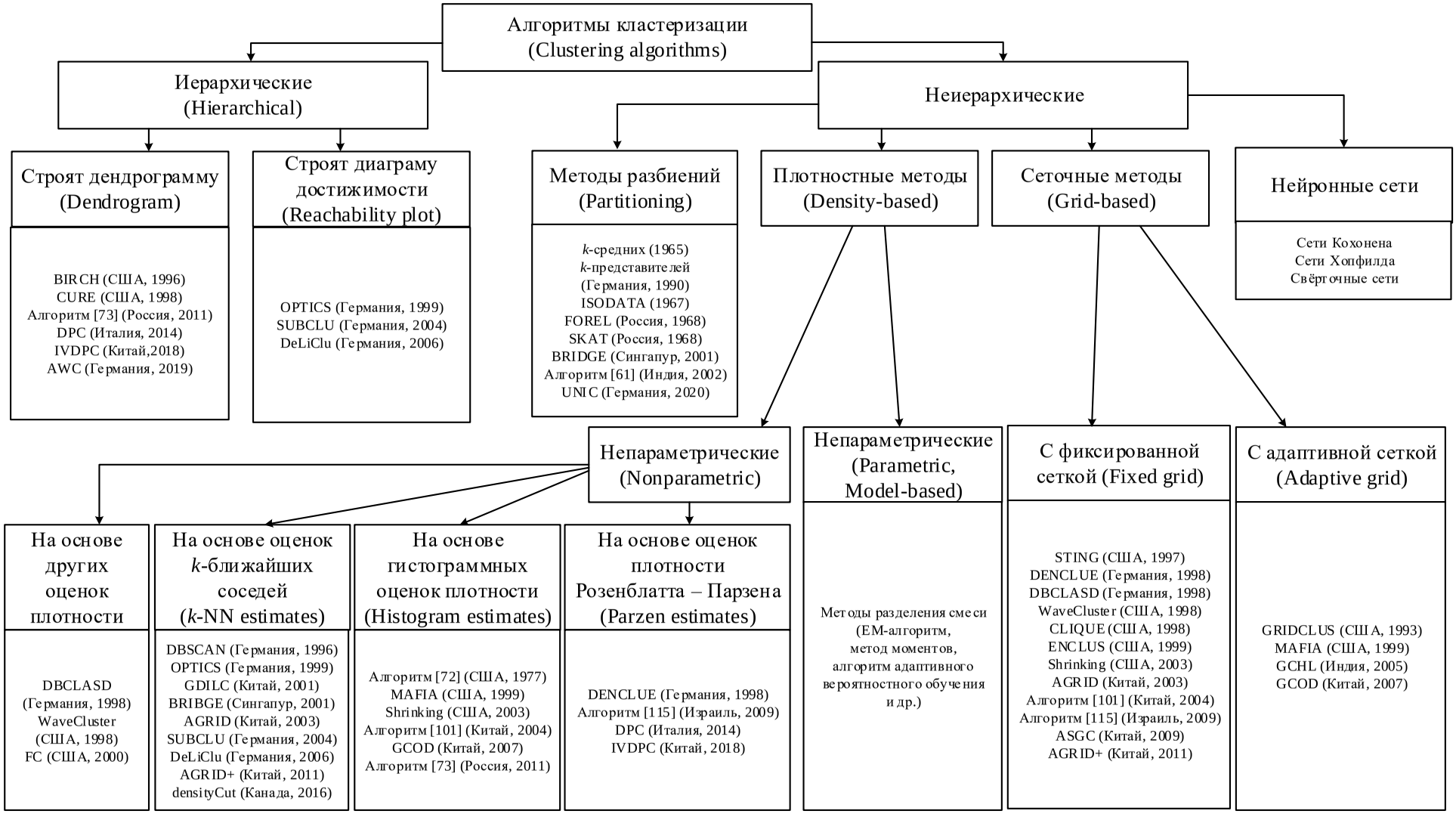


Рисунок 1.1 – Группы алгоритмов кластеризации

1.4.1. Иерархические алгоритмы

Иерархические алгоритмы [54-58] позволяют построить либо дерево кластеров, называемое дендрограммой (dendrogram), либо так называемую диаграмму достижимости (reachablity plot) [51]. Требуемый уровень детализации результата достигается за счёт отсечения построенной дендрограммы (или диаграммы достижимости) на определённом уровне с последующим разбиением исходного множества объектов на кластеры.

По способу построения дендрограммы алгоритмы можно разделить на агломеративные и разделительные. При использовании агломеративных (agglomerative) алгоритмов каждый объект считается одноэлементным кластером, после чего выполняется поэтапное объединение наиболее похожих кластеров. Для разделительных (divisive) алгоритмов, наоборот, вся выборка изначально считается одним кластером и на каждом шаге один из кластеров разбивается на два. И агломеративные, и разделительные алгоритмы просты в реализации, результат их выполнения не зависит от порядка ввода данных. К их недостаткам можно отнести высокую вычислительную сложность (порядка $O(N^3)$), которая не позволяет применять их для обработки спутниковых изображений. Кроме того, они не способны одновременно выделять кластеры разной структуры (формы, размера, плотности).

Для увеличения объёма данных, обрабатываемых произвольным агломеративным или разделительным алгоритмом кластеризации, можно использовать методы дробления [54] и повторного дробления [55]. Идея, лежащая в основе этих методов, заключается в разбиении всего множества данных на подмножества («дроби»), которые могут быть обработаны за приемлемое время, с последующим применением алгоритма кластеризацией к каждой «дроби». Затем из кластеров, полученных в результате обработки «дробей», формируется множество векторов-описаний (мета-объектов), которое впоследствии разбивается на группы с помощью того же алгоритма кластеризации. Одним из примеров алгоритма, использующего дробление, является BIRCH [56]. Основной недостаток алгоритма дробления заключается в том, что формирование мета-объектов может привести к ошибке

классификации, которую невозможно исправить на последующих этапах обработки. Алгоритм повторного дробления позволяет бороться с этим недостатком. Его суть заключается в многократном применении алгоритма дробления: построение «дробей» для последующей итерации осуществляется на основе результатов предыдущей.

Во многих агломеративных и разделительных алгоритмах для описания кластеров используются представители. Представителем кластера может служить как объект исходного множества, так и вектор средних значений (центр масс) кластера. Недостатком всех таких алгоритмов является неспособность выделять кластеры сложной формы (форма выделяемых кластеров жёстко детерминирована и зависит от используемой метрики). Один из способов сглаживания этого недостатка и основанный на нём алгоритм CURE описаны в [57]. В качестве представителя каждого кластера используется набор объектов, распределённых недалеко от границы. Для повышения разделимости близких кластеров набор представителей «сжимается» (каждый представитель немного сдвигается в направлении центроида соответствующего кластера). Такой подход позволяет разделять кластеры достаточно сложной формы (сложность зависит от количества представителей в описании кластера), но приводит к значительному росту времени обработки.

При обработке больших массивов данных построенная дендрограмма является очень сложной, её тяжело интерпретировать. В таких случаях для визуализации иерархической структуры кластеров целесообразней использовать диаграмму достижимости, на которой приведены расстояния достижимости для точек исходной выборки (в соответствии с введённым порядком). Согласно определению [51], расстояние достижимости для точки обратно пропорционально плотности в этой точке, поэтому «долины» на диаграмме достижимости соответствуют плотным областям пространства (кластерам), а вложенные «долины» – вложенным кластерам.

Диаграмма достижимости, в отличие от дендрограммы, не обладает древовидной структурой. Процесс её построения не может быть остановлен при достижении определённой степени раздробленности. Тем не менее, диаграмма достижимости позволяет восстановить иерархическую структуру кластеров. Поэтому алгоритмы,

строящие диаграмму достижимости, являясь иерархическими, не могут быть отнесены ни к агломеративным, ни к разделительным.

Несмотря на предложенные в последние годы модификации, алгоритмы построения дендрограммы «напрямую» не применимы для обработки спутниковых изображений ввиду высокой вычислительной сложности и невозможности одновременного выделения кластеров разной структуры. Однако их использование оправдано на заключительных этапах многоэтапных процедур (когда объём обрабатываемых данных становится небольшим), т.к. информация об иерархической вложенности кластеров существенно облегчает интерпретацию результатов сегментации.

1.4.2. Методы разбиений

Методы разбиений (partitioning methods) [49, 59-69] основаны на поэтапном улучшении некоторого начального разбиения исходного множества до получения оптимального значения заранее заданной заданной целевой функции. В качестве целевой функции часто используется сумма расстояний от объектов до центров кластеров, к которым они отнесены. Основным недостатком методов разбиений, использующих эту целевую функцию, является сильная зависимость формы выделяемых ими кластеров от используемой метрики. На практике это зачастую приводит к явным ошибкам кластеризации и/или чрезмерной раздробленности выделенных кластеров. Кроме того, методы разбиений не способны рассмотреть все возможные разбиения и есть вероятность обнаружения локального, а не глобального экстремума целевой функции. Для нахождения глобального экстремума необходимо рассмотреть $S(N, M)$ возможных разбиений исходной выборки на кластеры, где

$$S(N, M) = \frac{1}{M!}\sum_{i=1}^{M}(-1)^{M-i}\binom{M}{i}i^N.$$

Это одно из чисел Стирлинга второго рода, с ростом N оно очень быстро становится огромным.

Одним из первых методов разбиений является алгоритм k-средних (k-means) [59]. Его реализации включены практически во все пакеты программ, предназначенные для анализа спутниковых изображений. Алгоритм позволяет разбить исходную выборку на M кластеров (M – параметр, задаваемый пользователем). Полученные кластеры описываются векторами средних значений (центроидами). В процессе разбиения выполняется итеративная минимизация внутриклассовых расстояний. Соответствующая целевая функция выглядит следующим образом:

$$\sum_{j=1}^{M} \sum_{x^{(i)} \in C^{(j)}} \left\| x^{(i)} - c^{(j)} \right\|^2 \to \min,$$

где $c^{(j)}$ – центроид кластера $C^{(j)}$.

Вычислительная сложность алгоритма k-средних невысока (порядка $O(dMN)$), но он обладает несколькими недостатками. Несмотря на доказанную в [60] сходимость итеративного процесса, он не гарантирует нахождение глобального минимума целевой функции (оптимального разбиения). Поэтому для получения хорошего разбиения необходимо выбрать осмысленные начальные центроиды. Существуют эффективные методы выбора стартовых центроидов, гарантирующие получение качественных результатов кластеризации, но универсального эффективного метода нахождения оптимальных центроидов, не зависящего от структуры данных и специфики решаемой задачи, не может существовать даже теоретически [43]. Кроме того, алгоритм k-средних не способен выделять кластеры разной структуры. Для устранения этого недостатка в [61] предлагается разбивать исходную выборку на сферические кластеры, которые впоследствии используются для формирования итогового разбиения.

Ещё одним серьёзным недостатком алгоритма k-средних является необходимость задания числа кластеров, которое на практике чаще всего неизвестно и не существует эффективных методов его нахождения. Разработано несколько эвристических методов оценивания числа кластеров. Например, в алгоритме ISODATA [62] (наиболее распространённой модификации k-средних) число кла-

стеров может изменяться за счёт разбиения или объединения уже найденных в соответствии со значениями настраиваемых параметров. В итоге, задача определения числа классов сводится к настройке параметров алгоритма. Критерием разбиения служит диаметр кластера, а критерием объединения – расстояние между центроидами соседних кластеров. Алгоритм является итеративным, число кластеров, полученное на предыдущей итерации, используется как стартовое для инициализации последующей.

Помимо перечисленных недостатков, алгоритм k-средних чувствителен к выбросам в данных и «шуму», т.к. при вычислении центроида используются все точки кластера. В алгоритме ISODATA кластеры со слишком большим диаметром, которые с большой вероятностью содержат выбросы, в процессе обработки разбиваются, благодаря чему выбросы постепенно выделяются в отдельные кластеры. После этого кластеры, содержащие малое число точек, исключаются из рассмотрения. Поэтому алгоритм ISODATA не чувствителен к выбросам в данных. Реализации метода ISODATA включены во многие пакеты программ, предназначенные для обработки спутниковых изображений, поэтому исследователи часто используют его при решении практических задач. Однако для его успешного применения необходима кропотливая настройка входных параметров.

Алгоритм FOREL [63], как и алгоритм k-средних, позволяет минимизировать внутриклассовые расстояния, но в соответствии с ним на каждой итерации выделяется по одному кластеру. Для обнаружения кластера центр сферы фиксированного радиуса помещается в произвольную точку выборки. После этого центр итеративно перемещается в центр масс точек выборки, попавших в сферу, до достижения устойчивого состояния. Затем точки, попавшие в сферу, относятся в один кластер и исключаются из дальнейшего рассмотрения. Процедура выделения кластеров продолжается вплоть до отнесения всех точек выборки в соответствующие кластеры. Радиус сферы определяется итеративно, в зависимости от требуемого числа кластеров. Такой метод позволяет выделять кластеры сферической формы. Результаты выполнения алгоритма FOREL зависят от выбора начальных центров.

Для устранения этого недостатка в [49] описано несколько его модификаций. В алгоритме SKAT предлагается точки, отнесённые к кластерам, не исключать из рассмотрения. Это позволяет выделить кластеры, которые при использовании алгоритма FOREL сливаются с другими, и получать устойчивое разбиение. Алгоритм KOLAPS позволяет выделять сферические кластеры разного размера в присутствии «шума». В соответствии с ним кластер, содержащий малое число точек, выбрасывается из рассмотрения на последующих итерациях, но его точки помечаются как «шум». На заключительном этапе алгоритма для каждого выделенного кластера, начиная с наибольшего, ищется оптимальный радиус сферы, после чего точки, не попадающие в сферу оптимального радиуса, относятся к «шуму». Алгоритмы класса FOREL характеризуются слишком высокой трудоёмкостью для применения непосредственно к спутниковым изображениям и не позволяют выделять кластеры сложной структуры.

В отличие от k-средних, в алгоритме k-представителей (k-medoids) для описания кластеров используются объекты, взятые из исходной выборки (представители). Кластеры формируются путём отнесения каждой точки выборки к ближайшему представителю. Алгоритм k-представителей в наиболее общем виде описан в [64]. Он характеризуется высокой вычислительной сложностью (порядка $O(MN^2)$). Для обработки больших объёмов данных предложено несколько модификаций этого алгоритма [64, 65], в которых алгоритм k-представителей применяется к рабочей выборке (случайной выборке небольшого объёма), взятой из множества X. После этого каждый элемент исходной выборки относится к ближайшему элементу рабочей. Это позволяет существенно повысить производительность алгоритма, но при малом числе представителей некоторые кластеры могут быть упущены. Кроме того, в [66, 67] предложены модификации алгоритма для обработки данных высокой размерности.

В алгоритме FINDIT [68] используется мера схожести, учитывающая как расстояние между точками, так и степень информативности признаков. Алгоритм является модификацией метода k-представителей и позволяет обрабатывать данные

высокой размерности. Под степенью схожести точек понимается количество признаков, в которых точки различаются не более чем на заданное пороговое значение. В основе этой меры лежит предположение, что при кластеризации в многомерном пространстве признаков предпочтительней объединить точки, достаточно похожие по многим параметрам, чем точки, сильно похожие по нескольким признакам. Такой подход позволяет с легкостью исключить из рассмотрения «шум» и выбросы в данных.

Общим недостатком методов разбиения является сильная зависимость результата от значений настраиваемых параметров. Кроме того, они (за исключением алгоритма, предложенного в [61]) не способны выделять кластеры разной структуры (размер, форма и плотность выделяемых кластеров сильно зависят от используемой метрики). Несмотря на это, методы разбиений могут быть эффективно использованы в качестве отдельных этапов многоэтапных процедур сегментации спутниковых изображений.

1.4.3. Плотностные методы

Плотностные методы (density-based methods) рассматривают исходную непомеченную выборку как набор реализаций некоторого случайного вектора $\mathbf{x}$. Они разбивают объекты на кластеры на основе оценки многомерной плотности распределения $f(\mathbf{x})$. В данном случае под кластером понимается связная плотная область в пространстве признаков. Такое определение позволяет выделять кластеры сложной формы и кластеры разного размера. Однако вычисление оценок плотности распределения при обработке спутниковых изображений связано с неприемлемо высокой трудоёмкостью.

В зависимости от типа используемых оценок, плотностные алгоритмы подразделяются на параметрические (parametric, model-based), и непараметрические (nonparametric).

При параметрическом подходе предполагается, что функция $f(\mathbf{x})$ описывается заранее определённой вероятностной моделью с фиксированным набором настраиваемых параметров. Параметрические методы рассматривают $f(\mathbf{x})$ как

смесь, состоящую из M независимых распределений (чаще всего, гауссовских) с неизвестными параметрами. В этом случае задачу кластеризации можно решать как задачу разделения смеси: 1) с помощью алгоритма EM [70] оценить параметры моделей по непомеченным данным; 2) используя полученные параметры, построить разбиение изображения по методу максимального правдоподобия.

Алгоритмы, разработанные в рамках параметрического подхода, обладают рядом достоинств (слабая чувствительность к «шуму» и выбросам, независимость результата кластеризации от порядка ввода данных, способность выделять кластеры разной структуры и др.), однако для их корректного применения необходимы априорные сведения о параметрической структуре данных и количестве классов. Кроме того, параметрические алгоритмы характеризуются высокой вычислительной сложностью.

Оценки, используемые непараметрическими алгоритмами, строятся на основе анализа исходных данных и накладывают слабые ограничения на вид плотности распределения (непрерывность, ограниченность и т.п.). Благодаря этому непараметрические алгоритмы позволяют выделять кластеры сложной структуры. Однако для вычисления непараметрической оценки плотности в произвольной точке пространства признаков необходимо учесть вклад каждой точки исходной выборки, поэтому применение плотностных алгоритмов «напрямую» для обработки спутниковых изображений зачастую приводит к неприемлемым вычислительным затратам. Наиболее распространенными непараметрическими оценками плотности распределения являются гистограммная оценка (histogram estimation), оценка k-ближайших соседей (k-nearest neighbors estimation, k-NN estimation) и оценка плотности Розенблатта – Парзена (Parzen density estimation).

Для получения гистограммной оценки координатные оси пространства признаков разбиваются на интервалы, как правило, одинаковой длины (на основании этого пространство признаков разбивается на непересекающиеся ячейки), и подсчитываются частоты попадания значений векторов-признаков в полученные ячейки. При выполнении определённых условий [43, 71], гистограммная оценка является состоятельной.

Один из наиболее распространённых алгоритмов на основе гистограммной оценки плотности предложен в [72]. Он позволяет выделять кластеры сложной формы и разного размера, характеризующиеся одномодальным распределением, даже в присутствии «шума». В [73] предложена многоуровневая иерархическая процедура на основе этого метода, позволяющая выделять кластеры, характеризующиеся многомодальным распределением.

В основе оценки k-ближайших соседей лежит предположение, что вероятность принадлежности двух элементов выборки одному классу обратно пропорциональна расстоянию между ними («подобное – к подобному»). При использовании этой оценки непомеченный элемент выборки относится к тому же кластеру, что и большинство из k ближайших к нему помеченных элементов. Оценка k-ближайших соседей в чистом виде чувствительна к «шуму», поэтому расстояние, на котором соседние элементы влияют друг на друга, часто ограничивают пороговым значением. Эта оценка плотности является состоятельной, но смещённой.

Примером вычислительно эффективного алгоритма на основе оценки k-ближайших соседей является DBSCAN [74]. Для построения оценки плотности, на основе соседства точек вводятся понятия достижимости и связности. Под *ε-соседями* точки $x \in X$ понимается множество точек, расстояние до которых не превышает ε, т.е. $N_\varepsilon(x) = \{y \in X | D(x,y) \leq \varepsilon\}$. Точка y *достижима* из точки x, если существует последовательность точек $x_0 = x, x_1, \dots, x_{p-1}, x_p = y$, для которой выполнено:

$$x_{i+1} \in N_\varepsilon(x_i), i = 1, \dots, p-1; \qquad |N_\varepsilon(x_i)| \geq MinPts, i = 1, \dots, p-1.$$

Здесь значение $MinPts$ задаётся пользователем и регулирует порог отсечения «шума». Согласно второму условию, у точек, находящихся внутри кластера, должно быть не менее $MinPts$ ε-соседей. Такие точки называются «*ядрами*». Остальные точки разделяются на граничные (имеющие менее $MinPts$ ε-соседей, но достижимые из какого-либо «ядра») и шумовые. Две точки считаются *связными*, если существует «ядро», из которого они обе достижимы.

При такой постановке задачи, под *кластером* понимается максимальное связное подмножество множества X. Точки, не попавшие в какой-либо кластер (не принадлежащие ε-окрестности какого-либо «ядра»), относятся к «шуму».

К настоящему времени разработано достаточно много модификаций алгоритма DBSCAN. В [75] предложен обобщённый алгоритм, позволяющий обрабатывать объекты, которые представлены в пространстве признаков не только точками, но и, например, полигонами. В [76] предложена параллельная версия алгоритма для работы на высокопроизводительных вычислительных системах, а в [77] – способ потоковой обработки данных.

Для запуска DBSCAN необходимо задать два параметра, оптимальные значения которых определить достаточно сложно. Поэтому в [51] предложен алгоритм OPTICS, позволяющий упорядочить исходное множество и упростить процесс кластеризации. В соответствии с ним строится диаграмма достижимости, благодаря которой появляется возможность при фиксированном значении $MinPts$ обработать не только заданное значение ε, но и все $\varepsilon^* < \varepsilon$.

Для упорядочивания множества X для каждого его элемента вычисляется два параметра – «ядерное расстояние» (core distance, CD) и наименьшее из «расстояний достижимости» (reachability distance, RD):

$$CD(x) = \begin{cases} +\infty, & \text{если } |N_\varepsilon(x)| < MinPts, \\ MinPts\_dist(x), & \text{иначе;} \end{cases}$$

$$RD(x,y) = \begin{cases} +\infty, & \text{если } |N_\varepsilon(y)| < MinPts, \\ \max\{CD(y), D(x,y)\}, & \text{иначе;} \end{cases}$$

Здесь $MinPts\_dist(x)$ – расстояние от точки x до её $MinPts$-го соседа.

Проще говоря, $CD(x)$ – это наименьшее значение ε^*, при котором x является «ядром», а «расстояние достижимости» – значение ε^*, при котором x становится напрямую достижима из y. Соответственно, $RD(x)$ – наименьшее из «расстояний достижимости» для x – это значение ε^*, при котором x становится достижимой хотя бы из одного «ядра» (перестаёт быть шумовой). В зависимости от комбинации

этих параметров, при фиксированном значении ε^* точка x может быть как внутренней ($CD(x) \leq \varepsilon^*$), так и граничной ($CD(x) > \varepsilon^*$, $RD(x) \leq \varepsilon^*$) точкой кластера, а также являться шумовой ($CD(x) > \varepsilon^*$, $RD(x) > \varepsilon^*$).

Благодаря сортировке с помощью дополнительных полей, классификация выборки алгоритмом DBSCAN с параметрами $\varepsilon^* \leq \varepsilon$, $MinPts$ сводится к последовательному перебору упорядоченной выборки и присвоению каждому её элементу номера соответствующего кластера (отсечению диаграммы достижимости на нужном уровне и выделению на ней кластеров).

Экспериментально установлено [51], что OPTICS работает примерно в 1.6 раза медленнее, чем DBSCAN. Для запуска алгоритма OPTICS (как и DBSCAN) требуется два параметра – ε и $MinPts$. На данный момент предложено множество его модификаций (в том числе потоковая версия [78], позволяющая быстро пересчитывать кластеры при появлении новых точек), а также параллельная версия для обработки данных на многопроцессорных вычислительных системах [79].

Результат работы OPTICS зависит от параметра ε гораздо слабее, чем результат DBSCAN. Однако при слишком маленьком ε иерархическая структура классов может остаться незамеченной (вплоть до отнесения всей выборки в класс «шум»), а при слишком большом вычислительная сложность алгоритма становится неприемлемо высокой. Для решения этой проблемы в [80] предложен алгоритм DeLiClu, в соответствии с которым точки исходной выборки добавляются в диаграмму достижимости последовательно. Для ближайших соседей добавляемой точки методами вычислительной геометрии находятся $MinPts$ и $RD(x)$. Благодаря этому, для работы алгоритма DeLiClu необходим всего один параметр ($MinPts$). Основным достоинством алгоритма является то, что он при заданном значении $MinPts$ позволяет полностью восстановить иерархическую структуру кластеров.

Алгоритм SUBCLU [81] также разработан на основе DBSCAN. Он основан на следующем предположении: кластер, существующий в пространстве определённой размерности, существует и во всех его подпространствах. Основная идея алгоритма заключается в применении DBSCAN к проекциям исходной выборки на подпространства исходного пространства признаков. Алгоритм SUBCLU позволяет

выделить в подпространстве кластеры, которые выделил бы DBSCAN, применённый напрямую к этому подпространству. При этом SUBCLU обладает высокой производительностью.

В [82] предложен комбинированный трёхэтапный алгоритм BRIDGE. В соответствии с ним, сначала выборка разбивается на кластеры при помощи алгоритма k-средних (или BIRCH). Затем для выделения в каждом кластере «шума» используется DBSCAN, а на последнем этапе снова применяет k-средних к выборке уже без «шума». Такая схема позволяет сгладить некоторые недостатки обоих алгоритмов.

В алгоритме GDILC [83], как и в DBSCAN, плотность оценивается на основе соседства точек. Для этого в пространстве признаков вводится сеточная структура. С её помощью строятся изолинии плотности, по которым выделяются кластеры (области, окружённые изолиниями определённого уровня). Недостатком алгоритма GDILC является то, что он применим только к данным низкой размерности. В [84] на основе GDILC предложен алгоритм AGRID, позволяющий обрабатывать многомерные данные. Для оценивания плотности в точке, в AGRID используются не только точки, попавшие в соответствующую ячейку, но и точки из соседних с ней ячеек (соседними являются ячейки, имеющие общую границу размерности $k - 1$). С ростом размерности число ячеек растёт экспоненциально, поэтому в [85] предлагается разбивать первые несколько размерностей на большее число интервалов, а все остальные – на меньшее (алгоритм AGRID+). Если упорядочить признаки в соответствии со снижением информативности, то можно регулировать итоговое число ячеек без значительного снижения качества кластеризации. В дополнение к этому, для получения более точной оценки плотности в точке x вводится степень соседства ячеек, зависящая от размерности общей границы с клеткой, содержащей x. Для оценивания плотности используются все соседние клетки, причём вклад точек, попавших в них, прямо пропорционален степени соседства клеток.

В алгоритме densityCut [86] плотность оценивается в два этапа. Оценка плотности k-ближайших соседей, построенная на первом этапе, итеративно улучшается, что позволяет искусственно усилить контраст между областями с высокой плотностью и границами кластеров.

Оценка плотности Розенблатта – Парзена складывается из вкладов всех элементов выборки. Вклад каждого вектора признаков описывается функцией $\Phi(x)$. Формула для вычисления оценки плотности $\hat{f}_N(x, \Phi)$ в произвольной точке пространства признаков имеет вид [87, 88]:

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

где $\Phi(x)$ – колоколообразная функция (ядро), удовлетворяющая условиям [89]:

$$1)\ \Phi(x) \geq 0\ \forall x \in R^k, \quad 2)\ \sup_{x \in R^k} \Phi(x) < \infty,$$

$$3)\ \int_{R^k} \Phi(x)dx = 1, \quad 4)\ \lim_{\|x\| \to \infty} \|x\|^k \Phi(x) = 0.$$

Эти условия необходимы для того, чтобы оценка плотности Розенблатта – Парзена являлась несмещённой и состоятельной.

В алгоритме DENCLUE [90] используется оценка плотности Розенблатта – Парзена с гауссовским ядром:

$$\Phi_G\left(x, x^{(i)}\right) = \exp\left(-\frac{\left\|x - x^{(i)}\right\|^2}{2h^2}\right).$$

Точки, в которых достигаются локальные максимумы плотности распределения, находятся с помощью процедуры «среднего сдвига», предложенной в [91]. Процедурой «среднего сдвига» называются повторяющиеся движения от точки $y_0 = x \in R^k$ к $y_1 = m_h(y_0, \Phi)$, затем от y_1 к $m_h(y_1, \Phi)$ и т.д. до шага l, на котором $y_l = m_h(y_l, \Phi)$. Вектор

$$m_h(x, \Phi) = \frac{\sum_{i=1}^{N} x^{(i)} \Phi'\left(x, x^{(i)}\right)}{\sum_{i=1}^{N} \Phi'(x, x^{(i)})} - x$$

называется вектором «среднего сдвига». Его направление в точке x совпадает с градиентом оценки плотности $\hat{f}_N(x, \Phi)$ (направлением максимального роста плотности в этой точке). Более подробно процедура «среднего сдвига» будет рассмотрена в разделе 2.2.

Точка x называется *«точкой притяжения» для* y, если процедура «среднего сдвига», стартовавшая из y, сходится в x. При таком подходе одномодовый кластер $C(x^*)$ задаётся локальным максимумом x^* и является множеством точек, для которых x^* – «точка притяжения». Если значение плотности $\hat{f}_N(x^*, \Phi)$ меньше заданного порога ξ, то все элементы кластера $C(x^*)$ относятся к «шуму».

В соответствии с алгоритмом DENCLUE, пространство признаков перед кластеризацией разбивается на гиперкубические ячейки со стороной $2h$. Процедура «среднего сдвига» стартует только из точек, попавших в плотные (содержащие более ξ точек исходной выборки) и соседние с ними ячейки. Благодаря этому DENCLUE позволяет обрабатывать данные, содержащие «шум» и выбросы. Кроме того, он позволяет выделять кластеры разной структуры.

В алгоритме DPC [92] (и его модификации IVDPC [93]) в каждой точке $x \in X$ вычисляются выборочная оценка плотности распределения и минимальное расстояние до точки выборки с большей плотностью. Это позволяет построить дендрограмму, по которой определяется итоговое разбиение.

Альтернативный подход к оцениванию плотности и основанный на нём алгоритм DBCLASD предложены в [94]. Используя предположение, что расстояния между точками внутри кластера подчиняются равномерному распределению, кластер определяется как максимальное непустое подмножество множества X, имеющее равномерное распределение расстояний до ближайших соседей (с некоторым порогом доверия). Для определения равномерности распределения используется критерий χ^2. Граница кластера описывается полигоном, построенным с использованием сеточного подхода. Элементы выборки обрабатываются последовательно, поэтому результаты кластеризации зависят от порядка входных данных. Для смягчения этого недостатка предложены две модификации: 1) точки могут перемещаться между кластерами в процессе обработки и 2) точки, отнесённые к «шуму»,

рассматриваются повторно после формирования кластеров. К преимуществам алгоритма относится то, что он позволяет выделять кластеры сложной структуры без входных параметров, а к недостаткам – зависимость результатов обработки от порядка ввода данных.

Непараметрические алгоритмы строят разбиение на основе анализа исходных данных и не накладывают ограничений на структуру кластеров. Поэтому они позволяют выделять кластеры разной структуры при наличии «шума» и выбросов. К существенным недостаткам всех описанных в литературе непараметрических алгоритмов можно отнести высокую вычислительную сложность. Кроме того, кластеры, выделяемые алгоритмом «среднего сдвига», характеризуются излишней раздробленностью.

1.4.4. Сеточные методы

Сеточные методы (grid-based methods) основаны на введении сеточной структуры в пространстве признаков. Отличительной особенностью алгоритмов, относящихся к этой группе, является переход от обработки отдельных элементов выборки к обработке элементов сеточной структуры. В общем виде схема работы сеточного алгоритма кластеризации выглядит следующим образом [95].

Шаг 1. Построить разбиение пространства признаков на ячейки.

Шаг 2. Разбить ячейки на кластеры.

Шаг 3. Разбить исходную выборку на кластеры на основе кластеризации ячеек.

Результат выполнения сеточных алгоритмов не зависит от порядка ввода данных. При низкой вычислительной сложности (порядка $O(N) \ldots O(N \log N)$), они позволяют выделять кластеры сложной структуры. К недостаткам сеточных алгоритмов можно отнести сильную зависимость качества выделяемых кластеров от размера ячеек.

Выбор размера и способа построения сетки – достаточно сложная задача (часто размер сетки является параметром алгоритма). При слишком большом размере или неудачном расположении клеток происходит искусственное огрубление гра-

ниц кластеров, а в случае близких классов даже их объединение. Это может привести к грубым ошибкам кластеризации. При измельчении клеток происходит незначительное улучшение результата за счёт серьёзного роста времени обработки. Кроме того, слишком мелкие клетки часто приводят к чрезмерному дроблению кластеров.

Во многих сеточных алгоритмах в процессе классификации считается количество элементов исходной выборки, попавших в ячейку. Это значение используется для уменьшения объёма вычислений за счёт отсечения пустых (или практически пустых) клеток, а в некоторых алгоритмах и при объединении кластеров. Это позволяет считать некоторые сеточные алгоритмы плотностными (с гистограммной оценкой плотности распределения). В дальнейшем будем считать, что алгоритм, в котором плотности в ячейках сравниваются не только с пороговым значением, но и между собой, разработан в рамках комбинации плотностного и сеточного подходов.

Классическим примером сеточного алгоритма с фиксированной сеткой является STING [96]. Алгоритм разрабатывался для выполнения пространственных запросов к базам данных, но с его помощью можно осуществлять кластеризацию спутниковых снимков. В соответствии с алгоритмом STING в пространстве признаков вводится иерархическая сеточная структура. Построение сеточной структуры начинается с одной ячейки (содержащей всю исходную выборку), которая впоследствии разбивается на несколько более мелких (по умолчанию, четыре). Каждая ячейка описывается различными параметрами, как зависящими (вектор средних значений; дисперсия; минимальные и максимальные значения признаков; метка статистического распределения – «нормальное», «равномерное» или «отсутствует»), так и не зависящими от значений переменных (например, число попавших в клетку векторов). Параметры родительских ячеек могут быть вычислены только если известны параметры всех дочерних.

Кластеризация выполняется от корня дерева (или какого-то среднего уровня иерархии) вплоть до листьев. На каждом уровне дерева определяются ячейки, точки в которых подчиняются одному из законов распределения; на последующих

уровнях рассматриваются только их потомки. После рассмотрения всех ячеек кластеры, плотность которых выше заданного порога, выделяются методом поиска в ширину. Алгоритм STING обладает низкой вычислительной сложностью (порядка $O(N)$) и позволяет одновременно выделять кластеры разной структуры. Результат его выполнения не чувствителен к «шуму» (клетки с низкой плотностью в процессе обработки автоматически относятся к «шуму») и не зависит от порядка ввода данных. Недостатком алгоритма является то, что границы выделяемых кластеров сильно зависят от размера сетки и зачастую являются грубыми.

В алгоритме WaveCluster [97] изображение в пространстве признаков со введённой сеточной структурой рассматривается как цифровой сигнал (количество точек, попавших в ячейку сетки, считается значением сигнала в этой ячейке). При таком подходе граница кластера, на которой меняется распределение данных, соответствует высокочастотному сигналу, а внутренняя область кластера – низкочастотному сигналу с высокой амплитудой. Тогда для выделения кластеров могут быть использованы технологии обработки сигналов, например вейвлет-преобразование, позволяющее устранить «шум» и разделить части сигнала с различными частотами. При этом качество классификации можно регулировать количеством повторных вейвлет-преобразований. Алгоритм WaveCluster позволяет выделять кластеры сложной структуры в присутствии «шума» и не требует задания числа кластеров. Результаты выполнения алгоритма не зависят от порядка ввода данных. Кроме того он обладает низкой вычислительной сложностью (порядка $O(N)$) и позволяет обнаружить иерархическую вложенность кластеров.

В алгоритме FC [98] предлагается ввести в пространстве признаков серию сеток разного масштаба (ячейки каждой последующей сетки вдвое мельче предыдущей). Полученную сеточную структуру можно рассматривать как фрактал и вычислять для неё фрактальную размерность. На первом этапе кластеризации при помощи алгоритма «ближайший сосед» генерируется начальное разбиение. Основная идея второго этапа заключается в пошаговом распределении точек по исходным кластерам так, чтобы фрактальные размерности кластеров оставались неиз-

менными. После рассмотрения всей выборки кластеры с низкой фрактальной размерностью относятся к «шуму». Алгоритм FC обладает низкой трудоёмкостью (порядка $O(N)$) и нечувствителен к «шуму» и выбросам. Однако результат его выполнения сильно зависит от начального разбиения и порядка ввода данных.

Жёсткая зависимость границ выделяемых кластеров от размера ячеек является общим недостатком всех алгоритмов, использующих фиксированную сетку. Существует несколько путей его устранения. В [99] предложен алгоритм ASGC, в котором введённая сеточная структура после кластеризации сдвигается на половину размера ячейки в каждом направлении, после чего процесс кластеризации повторяется. Совместный анализ полученных разбиений позволяет повысить точность выделения границ кластеров.

В [100] предложен алгоритм, в котором при анализе учитываются характеристики не только рассматриваемой, но и соседних с ней клеток. Это позволяет предварительно «сжать» плотные области пространства признаков (в [100] используется гистограммная оценка плотности) и повысить разделимость кластеров. Для «сжатия» данных используется метод, аналогичный закону всемирного тяготения. Кроме того, вместо одной сетки алгоритм использует последовательность фиксированных сеток различного масштаба (подобно алгоритму FC), среди которых выбираются сетки, позволяющие лучше других описать структуру обрабатываемых данных. После выполнения кластеризации на каждой из отобранных сеток, среди результатов выбирается наилучший с точки зрения компактности кластеров. Описанный подход является слишком трудоёмким для применения непосредственно к мультиспектральным изображениям, но его можно комбинировать с другими методами для повышения качества результатов обработки.

В [101] для построения гистограммной оценки плотности предложено использовать не только точки, попавшие в ячейку, но и ближайшие точки соседних клеток. Алгоритм не позволяет полностью избавиться от «шума» и выбросов, но значительно снижает их влияние на результат за счёт искусственного повышения контраста между плотными и неплотными ячейками.

Для работы в многомерном пространстве признаков в рамках сеточного подхода были разработаны алгоритм CLIQUE [102] и его незначительная модификация ENCLUS [103]. В основу CLIQUE (как и в алгоритме SUBCLU [81]) положено предположение, что кластер, существующий в пространстве определённой размерности, гарантированно существует во всех его подпространствах меньшей размерности. Поэтому предлагается сначала выделять плотные интервалы (кластеры) во всех одномерных проекциях, а затем формировать из них кластеры более высоких размерностей. С ростом размерности кластеры, найденные на предыдущих итерациях, перестают быть плотными областями и отсекаются (ENCLUS отличается от CLIQUE только критерием отсечения). Здесь под кластером понимается максимальное множество связных плотных ячеек в пространстве признаков. Если рассматривать множество плотных ячеек в подпространстве как граф (ячейки связаны, если они имеют общую границу), то поиск кластеров аналогичен процессу выделения в графе связных подграфов.

Альтернативным путём устранения этого недостатка является использование адаптивной сетки (adaptive grid), т.е. разбиение пространства признаков на ячейки на основе анализа исходных данных. Типичным представителем алгоритмов с адаптивной сеткой является GRIDCLUS [104]. В соответствии с ним, разбиение пространства признаков на ячейки зависит от распределения исходных данных. Затем для каждой ячейки вычисляется относительная плотность. Формирование кластеров начинается с центров (более плотных ячеек), к которым постепенно присоединяются менее плотные, имеющие с ними общую границу. К достоинствам алгоритма GRIDCLUS можно отнести возможность обработки больших объёмов многомерных данных и высокое быстродействие, а также возможность выделять заранее неизвестное число кластеров, в том числе вложенных. К недостаткам – чувствительность к «шуму».

В алгоритме GCHL [105] используется сеточная структура из ячеек одинакового размера, которая вводится по мере поступления данных. При появлении объекта, не попадающего ни в одну из существующих ячеек, образуется новая ячейка.

Ячейки построенной таким способом сеточной структуры могут получиться достаточно сложной формы, т.к. область пересечения и лежащие в ней элементы исходной выборки относятся к ячейке, введённой раньше других. После введения сеточной структуры вычисляется относительная плотность каждой ячейки (отношение количества попавших в ячейку объектов исходной выборки к её относительному объёму). Если относительная плотность не превышает заданный порог, ячейка удаляется, а попавшие в неё точки считаются шумовыми. Затем из оставшихся ячеек строятся кластеры по той же схеме, что и в GRIDCLUS. К достоинствам алгоритма GCHL можно отнести невысокую трудоёмкость и возможность обработки больших массивов многомерных данных. Кроме того, алгоритм нечувствителен к «шуму» и позволяет выделять кластеры разной структуры. Основной его недостаток заключается в зависимости результата кластеризации от порядка ввода данных.

Сеточная структура, используемая в алгоритме GCOD [106], строится аналогичным образом: по мере появления данных в пространстве признаков формируются ячейки фиксированного размера, которые могут пересекаться. После этого, в соответствии с описанным критерием, из пересекающихся ячеек формируются кластеры. Изолированные ячейки, содержащие мало элементов исходной выборки и расположенные вдали от сформированных кластеров, считаются шумовыми. Такой подход позволяет выполнять обработку в многомерном пространстве признаков и выделять кластеры разной структуры в присутствии «шума». Основной недостаток алгоритма GCOD – зависимость результата кластеризации от порядка ввода данных.

Алгоритм MAFIA [107] похож на CLIQUE, но в нём вместо фиксированной сетки используется адаптивная. Это позволяет уменьшить число параметров алгоритма и повысить точность выделения границ кластеров. Ещё одно значительное отличие алгоритма MAFIA от CLIQUE заключается в том, что при выделении кластеров в подпространстве учитывается плотность ячеек. Эта особенность позволяет считать MAFIA как сеточным, так и плотностным алгоритмом. В [108, 109]

предложена схема параллельной реализации алгоритма MAFIA для выполнения его на высокопроизводительных вычислительных системах.

Переход от попиксельной обработки к анализу элементов сеточной структуры позволяет значительно снизить трудоёмкость алгоритмов кластеризации. Сеточные алгоритмы позволяют получить качественные результаты, но для этого необходима длительная и нетривиальная настройка параметров. Неправильный выбор параметров алгоритма зачастую приводит к снижению качества результата, особенно при необходимости точного разделения кластеров. Существует несколько способов устранения этого недостатка за счёт применения различных модификаций при формировании сетки (адаптивна сетка, подвижная сетка, одновременного использования нескольких сеток и др.), но все они приводят к снижению производительности.

1.4.5. Нейронные сети

Искусственная нейронная сеть [110, 111] представляют собой систему, в некоторой степени моделирующую работу биологических нейронов при анализе данных. Важной особенностью нейронных сетей является способность к обобщению накопленных знаний. Нейросетевой алгоритм представляет собой одно- или многослойную сеть, каждый слой которой состоит из множества вычислительных узлов (нейронов). У нейрона имеется несколько входных связей (синапсов), каждая со своим весом $w^{(i,j)}$, по которым он принимает поступающие сигналы. В зависимости от взвешенной суммы поступивших сигналов и заданной функции активации, нейрон может «возбудиться» и передать на выходные связи (аксоны) соответствующий сигнал.

Для кластеризации применяются нейронные сети трёх видов: сети Кохонена [53], сети Хопфилда [112, 113] и свёрточные сети. Сеть Кохонена содержит один скрытый слой нейронов. Число синапсов каждого нейрона совпадает с количеством переменных k, а количество нейронов – с требуемым числом кластеров M (меняя количество нейронов можно регулировать число кластеров в процессе обу-

чения). Обучение сети Кохонена начинается с задания небольших случайных значений весовой матрицы. В дальнейшем происходит модификация весов при подаче на вход векторов исходной выборки (процесс самоорганизации или самообучения сети). Для элемента $x \in X$ определяется ближайший к нему нейрон с номером l, который называется победителем:

$$l = \arg\min_i \sum_{j=1}^{d} \left(x^{(j)} - w^{(i,j)}\right)^2.$$

Это означает, вектор x отнесён к кластеру $C^{(l)}$ и на текущем шаге обучения будет изменяться только вес нейрона-победителя с номером l (принцип «победитель забирает всё»).

В более сложных сетях Кохонена (самоорганизующихся картах [53]) при обучении изменяются веса всех нейронов из окрестности победителя. В этом случае темп обучения обратно пропорционален расстоянию до победителя. Первоначально в окрестности каждого нейрона находятся все нейроны сети, но с каждым шагом обучения окрестность сужается. В конце этапа обучения изменяются только веса победителя.

Сети Хопфилда [112] состоят из $M \times N$ нейронов (M слоёв по N нейронов). Процесс обучения таких сетей заключается в многократной обработке исходного изображения до тех пор, пока весовые коэффициенты не стабилизируются. После обучения сегментация изображения выполняется на основе весовых коэффициентов сети. Сети Хопфилда характеризуются достаточно трудоёмким процессом обучения, но они позволяют выполнить параллельную реализацию некоторых методов разбиения (например, в [113] предложена реализация алгоритма k-средних). При кластеризации данных с использованием сетей Хопфилда (как и сетей Кохонена) элементы выборки обрабатываются последовательно, поэтому результат зависит от порядка ввода данных.

Процесс сегментации изображения свёрточной нейронной сетью можно условно разделить на три этапа. На первом этапе изображение «сворачивается»

(поэлементно умножается на матрицу (ядро), а результат аккумулируется в соответствующей позиции выходного изображения). На втором этапе выполняется классификация с использованием специального полносвязного слоя. Третий этап предназначен для формирования итоговой картосхемы. Важным достоинством свёрточных нейронных сетей является автоматическое выделение информативных признаков в процессе обучения.

Характерной особенностью всех нейросетевых алгоритмов является необходимость задания числа кластеров. Кроме того, им требуется процедура обучения.

1.5. Возможные пути дальнейшего развития

В таблице 1.1 представлены характеристики наиболее эффективных алгоритмов кластеризации. Учитывая перечисленные в разделе 1.3 особенности задачи, наиболее перспективными путями для разработки алгоритмов сегментации мультиспектральных изображений в условиях малой априорной информации являются:

1) разработка алгоритмов в рамках комбинации нескольких подходов;
2) построение многоэтапных процедур, позволяющих на каждом этапе эффективно использовать достоинства отдельных алгоритмов.

Разработка алгоритмов в рамках комбинации нескольких подходов позволяет объединить их достоинства и сгладить недостатки. Например, алгоритмы, разработанные в рамках комбинации плотностного и сеточного подходов (OPTICS [51], DeLiClu [80] и др.), характеризуются высокой (для плотностных методов) производительностью и качественным выделение границ кластеров, не свойственным сеточным методам.

Использование алгоритмов, разработанных в рамках разных подходов, для построения многоэтапных процедур позволяет применять каждый алгоритм в подходящих для его выполнения условиях. Кроме того, такой подход позволяет на завершающих этапах обработки улучшить результаты, которые получены вычислительно эффективными алгоритмами. Примеры такого рода алгоритмов приведены в [73, 82, 114].

Таблица 1.1 – Характеристики алгоритмов кластеризации

Название алгоритма	Число параметров	Вычислительная сложность	Обработка большого объёма данных	Результат не зависит от порядка ввода данных	Не требуется задание количества классов	Выделение кластеров сложной структуры	Выполнение в присутствии «шума» и выбросов	Обнаружение иерархической структуры кластеров	Число итераций определено заранее
BIRCH	1	$O(N)$	–	–	–	–	–	+	+
CURE	4	$O(N^2 \log N)$	–	+	–	+	+	+	+
k-средних	1	$O(dMN)$ (2)	+	+	–	–	–	–	–
ISODATA	6	$O(dMN)$ (2)	+	+	± (3)	–	+	–	+
k -представителей	1	$O(MN^2)$	–	+	–	–	–	–	–
DBSCAN	2	$O(N \log N)$	+	+	+	–	+	–	+
OPTICS	2	$O(N \log N)$	+	+	+	+	+	+	+
DeLiClu	нет	$O(N \log N)$	+	+	+	+	+	+	+
SUBCLU	2	$O(N \log N)$	+	+	+	+	+	–	+
BRIDGE	1	$O(N \log N)$	+	+	–	–	+	–	–
GDILC	2	$O(N)$	+	+	+	+	+	–	+
AGRID	2	$O(N)$	+	+	+	+	+	–	+
AGRID+	2	$O(dN)$	+	+	+	+	+	–	+
DBCLASD	нет	1.5-3 × DBSCAN	+	–	+	+	+	–	+
DENCLUE	2	$O(N \log N)$	+	–	+	+	+	–	+
GCOD	1	$O(dNK)$ (4)	+	–	+	+	+	–	+
Алгоритм [115]	2	$O(\widetilde{N}^2)$ (5)	+	+	+	+	+	–	+
STING	нет	$O(N)$	+	+	+	+	+	–	+
WaveCluster	нет	$O(N)$ при малом M	+	+	+	+	+	–	+
FC	2	$O(N)$	+	–	+	+	+	–	+
ASGC	3	$O(N + K)$ (4)	+	+	+	+	+	–	+
Алгоритм [100]	6	$O(N \log N)$ (2)	–	+	+	+	+	–	+
CLIQUE	2	$O(dN + M^d)$	+	+	+	+	+	–	+
ENCLUS	3	$O(dN)$	+	+	+	+	+	–	+
GRIDCLUS	3	$O(N)$	+	+	+	+	–	+	+
GCHL	2	$O(dN \log N)$	+	–	+	+	+	–	+
MAFIA	2	$O(dN + \text{const}^d)$	+	+	+	+	+	–	+

(1) В зависимости от сложности алгоритма поиска ближайших соседей
(2) Вычислительная сложность каждой итерации
(3) Необходимо задать лишь примерное число кластеров
(4) K – число ячеек сеточной структуры
(5) $\widetilde{N}$ – объём рабочей выборки

Процесс сегментации спутниковых изображений целесообразно разделить на три этапа: 1) «сжатие» данных с использованием алгоритмов с низкой вычисли-

тельной сложностью (например, сеточных); 2) выделение предварительных кластеров и «шума» с помощью плотностных алгоритмов или методов разбиений; 3) построение итогового разбиения. На последнем этапе целесообразно использовать иерархические алгоритмы, позволяющие представить результаты обработки в древовидной структуре, что значительно облегчает их интерпретацию.

Общей особенностью большинства алгоритмов кластеризации является сильная зависимость результата от значений настраиваемых параметров. Для устранения этого недостатка в последние годы активно развивается ансамблевый подход [15]. Алгоритмы, разработанные на его основе, характеризуются более высокой устойчивостью к изменению параметров [116, 117]. Более подробно ансамблевый подход будет рассмотрен в разделе 3.1.

Таким образом, несмотря на большое количество разработанных алгоритмов, задача разработки эффективных алгоритмов кластеризации для сегментации спутниковых изображений до сих пор остаётся актуальной.

Выводы по главе

1. Приведена содержательная постановка задачи кластеризации данных. Исходя из характерных особенностей задачи, сформулированы требования к алгоритмам кластеризации для сегментации спутниковых изображений.
2. Выполнен анализ известных алгоритмов кластеризации применительно к задаче сегментации мультиспектральных изображений.
3. Перечислены наиболее перспективные пути для разработки алгоритмов сегментации мультиспектральных изображений в условиях малой априорной информации.

ГЛАВА 2. НЕПАРАМЕТРИЧЕСКИЙ АЛГОРИТМ КЛАСТЕРИЗАЦИИ MEANSC (MEAN SHIFT CLASSIFIER)

2.1. Формальная постановка задачи кластеризации в рамках вероятностно-статистического подхода

Пусть имеется некоторое множество объектов Ω. Каждый объект $\omega^{(i)} \in \Omega$, $(i = \overline{1,N})$ описывается в системе переменных (признаков) $X_1, \dots, X_k$ набором из k числовых измерений (признаков), который можно рассматривать как вектор-столбец $x^{(i)} = \left(x_1^{(i)}, \dots, x_k^{(i)}\right)^T \in R^k$, где $x_j^{(i)}$ – значение признака X_j для объекта $\omega^{(i)}$.

Будем предполагать, что векторы $x^{(i)}$ по своей природе являются случайными. Тогда задачу кластеризации можно рассматривать в рамках вероятностно-статистического подхода: всё множество векторов $X = \left\{x^{(i)}, i = \overline{1,N}\right\}$ является набором реализаций некоторого случайного вектора **x**, плотность распределения которого $f(x)$, $x \in R^k$ неизвестна и нет никакой априорной информации о её параметрическом виде (что она может быть записана в аналитической форме, зависящей от фиксированного заранее известного числа параметров). Введём следующие определения.

Определение 2.1. *Пусть в точке $x^* \in R^k$ достигается локальный максимум функции плотности $f(x)$. Будем считать, что точка $x \in R^k$ связна с x^*, если существует отрезок кривой $p_x(t) \in R^k$, $t \in [0,1]$:*

$$p_x(0) = x, \;\; p_x(1) = x^*, \;\; (t_1 - t_2)\left(f\left(p_x(t_1)\right) - f\left(p_x(t_2)\right)\right) \geq 0 \;\forall t_1, t_2 \in [0,1].$$

Определение 2.2. *Кластером, определяемым локальным максимумом плотности x^*, назовём непустое множество $Q(x^*) \subseteq R^k$, каждая точка которого связна с x^*. Если $f(x^*) < \varepsilon$, то кластер $Q(x^*)$ будем считать «шумом». Здесь $\varepsilon > 0$ – порог «шума».*

Тогда задача кластеризации заключается в том, чтобы на основе анализа набора векторов-признаков X разбить пространство R^k на заранее неизвестное число кластеров, определяемых локальными максимумами плотности, и «шум».

2.2. Оценка плотности Розенблатта – Парзена и процедура «среднего сдвига»

Для оценивания плотности распределения $f(x)$ в произвольной точке $x \in R^k$ пространства признаков по множеству X воспользуемся непараметрической оценкой Розенблатта – Парзена, вычисляемой по формуле [87, 88]

$$\hat{f}_N(x, \Phi) = \frac{1}{Nh^k} \sum_{i=1}^{N} \Phi\left(\frac{x - x^{(i)}}{h}\right). \quad (2.1)$$

Здесь h – параметр сглаживания; $\Phi(x)$ – ограниченная радиально симметричная функция (ядро), следующим удовлетворяющая условиям сходимости [87, 118]

$$\sup_{x \in R^k} |\Phi(x)| < \infty, \quad \int_{R^k} |\Phi(x)| dx < \infty,$$
$$\lim_{\|x\| \to \infty} \|x\| \Phi(x) = 0, \quad \int_{R^k} \Phi(x) dx = 1, \quad (2.2)$$

где символ $\|\cdot\|$ обозначает норму вектора. Значение параметра сглаживания h зависит от N, что позволяет адаптивно подстраиваться под структуру данных. При выполнении следующих условий

$$\lim_{N \to \infty} h(N) = 0, \quad \lim_{N \to \infty} Nh(N) = \infty, \quad \lim_{N \to \infty} Nh^{2k}(N) = \infty$$

построенная оценка является асимптотически несмещённой и состоятельной.

При обработке мультиспектральных изображений процедура вычисления ядерной оценки плотности (2.1) является трудоёмкой. Для уменьшения вычислительной сложности будем использовать финитные радиально-симметричные ядра вида

$$\Phi(x) = c_\phi \phi(\|x\|^2),$$

где $\phi(x)$ – неотрицательная скалярная функция, часто называемая «профилем» ядра $\Phi(x)$ [115, 119]; $c_\phi > 0$ – нормировочная константа, необходимая для выполнения условий (2.2). Заметим, что для построения финитного радиально-симметричного ядра достаточно задать функцию $\phi(x)$ на интервале $x \geq 0$.

Для поиска локальных мод плотности функции $f(x)$ воспользуемся процедурой «взбирания по градиенту». При использовании ядра с дифференцируемым профилем, градиент плотности распределения можно оценить через градиент оценки $\hat{f}_N(x, \Phi)$ следующим образом

$$\widehat{\nabla} f(x) = \nabla \hat{f}_N(x, \Phi) = \frac{1}{Nh^k} \sum_{i=1}^{N} \nabla \Phi \left(\frac{x - x^{(i)}}{h} \right) = \frac{c_\phi}{Nh^k} \sum_{i=1}^{N} \nabla \phi \left(\left\| \frac{x - x^{(i)}}{h} \right\|^2 \right) =$$

$$= \frac{2c_\phi}{Nh^{k+2}} \sum_{i=1}^{N} \left(x - x^{(i)} \right) \phi' \left(\left\| \frac{x - x^{(i)}}{h} \right\|^2 \right) \quad (2.3)$$

Предполагая, что производная функции $\phi(x)$ определена для всех $x \geq 0$, за исключением конечного множества точек, обозначим

$$\psi(x) = -\phi'(x).$$

Используя $\psi(x)$ в качестве профиля, определим ядро

$$\Psi(x) = c_\psi \psi(\|x\|^2)$$

и соответствующую ему оценку плотности

$$\hat{f}_N(x, \Psi) = \frac{1}{Nh^k} \sum_{i=1}^{N} \Psi \left(\left\| \frac{x - x^{(i)}}{h} \right\|^2 \right) = \frac{c_\psi}{Nh^k} \sum_{i=1}^{N} \psi \left(\left\| \frac{x - x^{(i)}}{h} \right\|^2 \right).$$

Здесь $c_\psi > 0$ – нормировочная константа.

Подставляя $\psi(x)$ в (2.3), получаем

$$\widehat{\nabla} f_N(x) = \frac{2c_\phi}{Nh^{k+2}} \sum_{i=1}^{N} \left(x^{(i)} - x \right) \psi \left(\left\| \frac{x - x^{(i)}}{h} \right\|^2 \right) =$$

$$= \frac{2c_\phi}{Nh^{k+2}} \left[\sum_{i=1}^{N} \psi \left(\left\| \frac{x - x^{(i)}}{h} \right\|^2 \right) \right] \left[\frac{\sum_{i=1}^{N} x^{(i)} \psi \left(\left\| \frac{x - x^{(i)}}{h} \right\|^2 \right)}{\sum_{i=1}^{N} \psi \left(\left\| \frac{x - x^{(i)}}{h} \right\|^2 \right)} - x \right]. \quad (2.4)$$

Считаем, что $\sum_{i=1}^{N} \psi \left(\left\| \frac{x - x^{(i)}}{h} \right\|^2 \right) > 0$ (это условие выполнено для всех радиально-симметричных ядер, используемых на практике). Первый множитель полученного выражения пропорционален оценке плотности $\hat{f}_N(x, \Psi)$, второй называется вектором «среднего сдвига»

$$m_h(x, \Psi) = \frac{\sum_{i=1}^{N} x^{(i)} \psi \left(\left\| \frac{x - x^{(i)}}{h} \right\|^2 \right)}{\sum_{i=1}^{N} \psi \left(\left\| \frac{x - x^{(i)}}{h} \right\|^2 \right)} - x.$$

Подставляя $\hat{f}_N(x,\Psi)$ и $m_h(x,\Psi)$ в (2.4), получаем

$$\widehat{\nabla} f_N(x,\Phi) = \hat{f}_N(x,\Psi)\frac{2c_\phi}{h^2 c_\psi} m_h(x,\Psi).$$

Иначе говоря,

$$m_h(x,\Psi) = \frac{h^2 c_\psi}{2c_\phi}\frac{\widehat{\nabla} f_N(x,\Phi)}{\hat{f}_N(x,\Psi)}. \quad (2.5)$$

Выражение (2.5) показывает, что вектор «среднего сдвига» $m_h(x,\Psi)$, вычисленный с использованием ядра Ψ, сонаправлен с оценкой градиента плотности $f_N(x,\Phi)$, построенной на основе ядра Ф. Это является обобщением свойства вектора «среднего сдвига», описанного в [14] и [30].

Следовательно, вектор «среднего сдвига» указывает направление максимального роста плотности распределения и позволяет определить кратчайший путь, ведущий к локальному максимуму функции $\hat{f}_N(x,\Phi)$. Стоит заметить, что величина «среднего сдвига», оставаясь большой в областях с низкой плотностью, постепенно уменьшается по мере приближения к моде [91].

Определение 2.3. *Итерационную процедуру, заключающуюся в переходе от* $x \in R^k$ *к* $x_1 = x + m_h(x,\Psi)$*, затем от* x_1 *к* $x_2 = x_1 + m_h(x_1,\Psi)$ *и т.д. до точки* x^* *для которой* $m_h(x^*,\Psi) = 0$*, называют процедурой «среднего сдвига». Путь, пройденный от точки* x *до локальной моды плотности* x^**, будем называть траекторией «среднего сдвига» и обозначать* $\hat{p}_x(t) = [x, x_1, \dots, x^*]$.

Оптимальным ядром в среднеквадратичном смысле является ядро Епанечникова [118]

$$\Phi_{\mathrm{E}}(\|x\|) = \begin{cases} \frac{1}{2}V_k^{-1}(k+2)(1-\|x\|^2), & \text{если } \|x\| < 1, \\ 0, & \text{иначе,} \end{cases}$$

где V_k – объём единичного k-мерного шара. Для этого ядра вектор «среднего сдвига» вычисляется по значительно более простой формуле [120]

$$m_h(x,\Psi_{\mathrm{E}}) = \frac{1}{N_x}\sum_{x^{(i)}\in V(x)} x^{(i)} - x,$$

где $V(x)$ – k-мерный шар радиуса h с центром в точке x; N_x – число точек, попавших в $V(x)$. Поэтому ядро Епанечникова иногда называют «тенью» равномерного ядра [121].

Пусть $y = \{y_i \in R^k,\ i = 0,1,\dots\}$ – траектория «среднего сдвига» для точки y_0, которая является оценкой кривой $p_{y_0}(t)$ из определения 2.1. Согласно построению,

$$y_{i+1} = y_i + \frac{h^2 c_\psi}{2c_\phi} \frac{\widehat{\nabla} f_N(y_i, \Phi)}{\hat{f}_N(y_i, \Psi)}.$$

Тогда справедлива следующая теорема (см. [119, 122]).

Теорема. *Пусть ядро* Ф *имеет выпуклый монотонно убывающий профиль* ϕ. *Тогда последовательность* $y = \{y_i \in R^k,\ i = 0,1,\dots\}$ *сходится. Кроме того, последовательность* $\hat{f} = \{\hat{f}_i = \hat{f}_N(y_i, \Phi), i = 0,1,\dots\}$ *монотонно возрастает и сходится к локальному максимуму оценки плотности* $\hat{f}_N(x, \Phi)$.

Процедура «среднего сдвига» является достаточно трудоёмкой, впервые её применение для сегментации изображений описано в [119]. Основная сложность заключается в необходимости многократного вычисления расстояния между точками в многомерном пространстве и векторов «среднего сдвига». Для решения этой проблемы в [18] предложено вместо исходной выборки при оценке плотности распределения использовать так называемую рабочую выборку. Экспериментально показана возможность применения этого метода для обработки (в том числе, сегментации) изображений; продемонстрировано, что даже при значительном уменьшении выборки (рабочая выборка в 1024 раза меньше исходной), качество результатов остаётся достаточно высоким, а время обработки многократно уменьшается.

Главная проблема при уменьшении выборки заключается в необходимости формирования представительной рабочей выборки. Однако ни один из предложенных в литературе методов решения этой проблемы не позволяет получить рабочую выборку, которая гарантированно содержит объекты из всех классов, присутствующих на изображении.

Альтернативой уменьшению выборки может служить уменьшение необходимого числа итераций «среднего сдвига». Для этого можно адаптивно подстраивать

параметр h под структуру данных [123, 124] или вычислять взвешенный вектор «среднего сдвига» [120].

В диссертационной работе рабочая выборка используется только в качестве набора векторов для запуска процедуры «среднего сдвига». Для её формирования предложена процедура, основанная на сеточном подходе к кластеризации данных. Она заключается во введении сеточной структуры в пространстве признаков с последующим выбором представителей от каждой ячейки, содержащей достаточно большое число точек исходной выборки. Это позволяет, с одной стороны, построить рабочую выборку, гарантированно содержащую представителей из всех классов, присутствующих на изображении. С другой стороны, он позволяет избавиться от «шума» и многократно уменьшить вычислительную сложность алгоритма за счёт сокращения числа точек, к которым необходимо применить процедуру «среднего сдвига». Кроме того, он не ведёт к потере точности при вычислении локальных мод плотности [18].

2.3. Выбор параметра сглаживания

Проблема выбора значения параметра сглаживания h является одной из основных при использовании ядерных оценок плотности. При слишком маленьком значении h появляются локальные максимумы функции $\hat{f}_N(x, \Phi)$, не соответствующие модам функции $f(x)$, а при слишком большом оценка получается чрезмерно сглаженной, что усложняет разделение близких кластеров. Существует три основных стратегии выбора значения параметра сглаживания: одинаковым во всём пространстве признаков $(h = const)$ [125], постоянным внутри каждого кластера $(h \in \{h^{(1)}, \dots, h^{(M)}\})$ [126] или в зависимости от точки выборки $(h = h(x^{(i)}), x^{(i)} \in X)$ [123, 127, 128]. Значение h может также зависеть от точки, в которой выполняется оценка плотности $(h = h(x), x \in R^k)$, однако построенная таким образом функция $\hat{f}_N(x, \Phi)$, строго говоря, не является оценкой плотности распределения, т.к. для неё не всегда возможно подобрать нормировочный коэффициент [124].

Использование адаптивного значения h позволяет более точно оценить плотность распределения, однако это приводит к значительному увеличению трудоёмкости. К настоящему времени предложено четыре основных способа вычисления значения параметра сглаживания [119], которое является оптимальным в соответствии с различными критериями.

1. *Статистически оптимальное:* выбирается значение h, позволяющее получить компромисс между смещённостью и дисперсией построенной оценки. В итоговой формуле [129, с. 85] присутствует вторая производная функции плотности распределения. При использовании гауссовского ядра, её можно вычислить (с помощью подстановочного правила [130]) или оценить [131].
2. *Оптимальное в смысле стабильности разбиения:* строится зависимость итогового числа кластеров от значения h; оптимальным считается середина самого широкого интервала, внутри которого число кластеров не изменяется [89].
3. *Оптимальное в смысле качества разбиения:* оптимальным считается значение параметра сглаживания, при котором достигается наибольшее значение функционала, характеризующего качество формируемых кластеров. Характеристика качества разбиения может зависеть от расстояний внутри кластеров и между ними [64], от степени изолированности кластеров [132] и др.
4. *Оптимальное для решаемой задачи:* значение h задаётся пользователем или выбирается исходя из анализа имеющейся априорной информации.

Исходя из вышеизложенного, все методы вычисления оптимального значения параметра сглаживания, предложенные в литературе, являются слишком трудоёмкими для применения их в задаче сегментации изображений. Поэтому в рамках диссертационной работы используется фиксированное значение h, задаваемое пользователем.

2.4. Предлагаемый алгоритм MeanSC

Предлагаемый алгоритм сегментации мультиспектральных изображений MeanSC (Mean Shift Classifier) разработан в рамках плотностного подхода на основе непараметрической оценки Розенблатта – Парзена с ядром Епанечникова. Для быстрого нахождения локальных мод плотности используется процедура «среднего сдвига».

Алгоритмы с ядерной оценкой плотности распределения характеризуются очень высокой вычислительной сложностью. В ходе диссертационной работы предложено несколько оптимизаций, основанных на специфике мультиспектральных изображений (см. раздел 1.3).

Основной особенностью мультиспектральных спутниковых изображений, а также главной проблемой при их сегментации, является большой объём исходных данных, который невозможно обработать классическими плотностными алгоритмами. Поэтому в пространстве признаков вводится сеточная структура с размером ячейки $2h$ (здесь h – параметр сглаживания в формуле (2.1)). Учитывая финитность используемого ядра Епанечникова, при вычислении оценки плотности и векторов «среднего сдвига» в произвольной точке $x \in R^k$ достаточно использовать только векторы, находящиеся на расстоянии не более h от x. Благодаря выбранному размеру ячейки, все такие векторы расположены в ячейках, соседних с ячейкой, в которую попала точка x (см. двумерный пример на рисунке 2.1). Для нахождения всех векторов, удалённых от x не более, чем на h, достаточно рассмотреть ячейки, с которыми пересекается сфера радиуса h с центром в точке x (всего четыре ячейки, выделенных серой заливкой). Таким образом, использование сеточной структуры позволяет многократно уменьшить необходимый перебор и снизить трудоёмкость алгоритма без снижения качества разбиения.

Второй характерной особенностью мультиспектральных изображений, позволяющей повысить быстродействие предлагаемого алгоритма, является ограниченность диапазонов изменения значений спектральных признаков, обусловленная фиксированным числом уровнем квантования выходного сигнала съёмочной

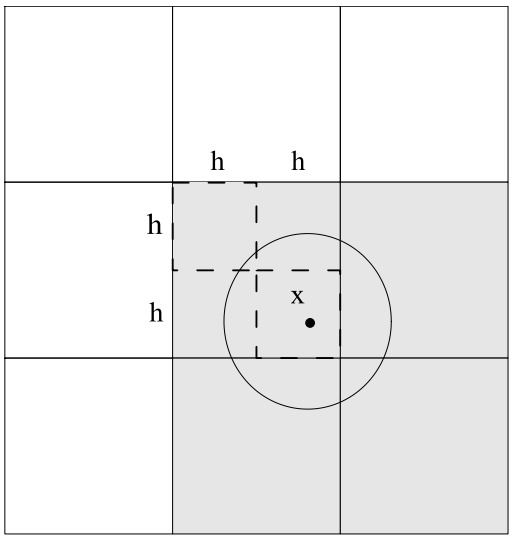


Рисунок 2.1 – Пример выбора соседних ячеек сеточной структуры в двумерном пространстве признаков

аппаратуры. Эта особенность позволяет ввести сеточную структуру в пространстве признаков до начала обработки, тем самым избавляясь от дополнительного просмотра изображения.

Третья особенность мультиспектральных изображений является следствием первых двух и заключается в высокой степени повторяемости значений признаков. Это позволяет, используя предварительное хеширование (объединяя все одинаковые вектора исходной выборки в один с весом, равным суммарному весу всех таких векторов), в несколько раз уменьшить объём исходной выборки без потери содержащейся в ней информации. Эксперименты на четырёх каналах мультиспектральных изображений (красный, зелёный, синий и ближний инфракрасный), показали уменьшение выборки от 7 раз на снимке QuickBird, до 61 раза на некоторых снимках Landsat-7.

Разбиение, получаемое при помощи процедуры «среднего сдвига» с фиксированным значением параметра h, характеризуется излишней раздробленностью, поэтому на последнем шаге предлагаемого алгоритма выполняется объединение кластеров, центры которых можно связать отрезком кривой, не содержащим точек с низкой плотностью («провалов» плотности). Здесь под центром кластера понимается мода плотности, которой он определяется (см. определение 2.2). Наличие «провала» плотности на отрезке проверяется с помощью оригинального критерия, который будет описан в определении 2.6.

Из финитности ядра Епанечникова, следует, что любая кривая, соединяющая центры кластеров, расстояние между которыми превышает $2h$ (согласно метрике «ближайший сосед» для множеств), обязательно содержит точку с нулевой плотностью. Поэтому на последнем этапе алгоритма достаточно проверить пары кластеров, расстояние между которыми не превышает $2h$. Для нахождения таких пар формируются описания всех найденных кластеров элементами сеточной структуры по следующему правилу: клетка относится к описанию кластера, если в неё попадает хотя бы одна точка из этого кластера. Заметим, что точки, попавшие в одну ячейку, могут быть отнесены к разным кластерам, поэтому описания кластеров могут пересекаться. При таком подходе, на последнем этапе алгоритма достаточно проверить пары кластеров, клетки из описаний которых пересекаются или являются смежными. Это позволяет значительно снизить трудоёмкость последнего этапа алгоритма, особенно при выделении разреженных кластеров. Из определения 2.2 следует, что любую точку выборки можно соединить с центром кластера, к которому она отнесена, отрезком кривой, вдоль которого значение плотности распределения монотонно возрастает (траекторией «среднего сдвига»). Поэтому искомая точка с наименьшим значением плотности гарантированно находится на отрезке кривой, соединяющем точки из разных кластеров.

Последняя оптимизация касается распределения точек исходной выборки по кластерам, сформированным из точек рабочей выборки. В [115] предложено отнести каждый элемент исходной выборки к тому кластеру, к которому отнесён ближайший к нему элемент рабочей выборки. В ходе выполнения диссертационной работы предложен другой метод, позволяющий снизить негативный эффект от использования рабочей выборки вместо исходной при «среднем сдвиге».

Предложенный метод заключается в следующем. Оригинальная процедура «среднего сдвига», описанная в [30], начинается из произвольной точки выборки и заключается в пошаговом перемещении центра из текущей точки в точку с наибольшей плотностью среди всех точек окрестности, вплоть до достижения локального максимума плотности. Процедура, описанная в [115], отличается от неё только тем, что на первом шаге обработки каждой точки исходной выборки центр

совмещается с ближайшим к ней представителем, включённым в рабочую выборку. Очевидно, что ближайший элемент рабочей выборки не всегда является точкой с наибольшей плотностью в окрестности обрабатываемого элемента исходной выборки. Часто это приводит к ошибкам на границах кластеров. Предложенный метод позволяет сгладить этот недостаток. Для распределения элементов исходной выборки по кластерам используется не только рабочая выборка, но и все точки, в которых вычислялось значение плотности при применении процедуры «среднего сдвига» к рабочей выборке (все точки всех построенных траекторий «среднего сдвига»). То есть элемент исходной выборки относится к тому же кластеру, что и точка с максимальной плотностью среди всех точек его окрестности, для которых было вычислено значение плотности. Это позволяет без дополнительных вычислений многократно увеличить объём выборки, используемой для распределения точек по кластерам.

На практике это достигается следующим образом. Для каждой точки исходной выборки запоминается набольшее из значений оценки плотности, вычисленных в её окрестности радиуса h. По завершении каждого шага процедуры «среднего сдвига», для всех точек исходной выборки, попавших в окрестность нового центра, выполняется повторное распределение по кластерам. Для ускорения этого процесса используется введённая сеточная структура. Это позволяет избежать хранения всех траекторий «среднего сдвига».

Описанные оптимизации, совместно с вычислением частичных сумм (везде, где это допустимо) и другими незначительными улучшениями, позволяют уменьшить время обработки изображения на два, а в некоторых случаях даже на три порядка.

Использование оценки плотности Розенблатта – Парзена и процедуры «среднего сдвига» позволяет модифицировать определения 2.1 и 2.2 следующим образом.

Определение 2.4. *Пусть в точке $x^* \in R^k$ достигается локальный максимум оценки плотности $\hat{f}_N(x)$. Тогда точка $x \in R^k$ связна с x^*, если процедура «среднего сдвига», стартовавшая из x, сходится к x^*. В дальнейшем верхним индексом «*»*

будем обозначать моду плотности, к которой сходится процедура среднего сдвига, стартовавшая из точки пространства.

Определение 2.5. *Кластером, определяемым локальным максимумом* x^* *(*$\hat{f}_N(x^*) \geq \varepsilon$*), назовём непустое подмножество точек* $Q(x^*) \subseteq X$*, связных с* x^**. Если* $\hat{f}_N(x^*) < \varepsilon$*, то кластер* $Q(x^*)$ *будем считать «шумом». Здесь* $\varepsilon \geq 0$ *– порог «шума».*

Разбиение исходной выборки на множество кластеров $\mathbb{C}$ в соответствии с определением 2.5 приводит к чрезмерно раздробленным результатам. Поэтому в ходе диссертационной работы предложен следующий критерий выделения многомодовых кластеров.

Определение 2.6. *Многомодовым кластером назовем непустое подмножество* $C \subseteq X$*, удовлетворяющее условиям:*

1) $\forall x \in C$ *выполнено* $Q(x^*) \in \mathbb{C}$*;*
2) $\forall x_1, x_2 \in C$ *существует кривая* $P \subset R^k$*, соединяющая* x_1^* *и* x_2^**, вдоль которой*

$$1 - \frac{\hat{f}_N(x)}{min\left(\hat{f}_N(x_1^*), \hat{f}_N(x_2^*)\right)} \leq T.$$

Здесь $T \in [0,1]$ *– параметр, отвечающий за уровень детализации результата.*

В соответствии со введёнными определениями, алгоритм MeanSC(m,ε,T) можно записать в виде следующей последовательности шагов.

Шаг 1. Формируем клеточную структуру данных в пространстве признаков. Для этого разбиваем пространство признаков $[0, K_1 - 1]_1 \times \ldots \times [0, K_k - 1]_k$ на гиперкубические клетки со стороной $2h$, где h – параметр сглаживания для оценки плотности $\hat{f}_N(x)$, вычисляемый по формуле

$$h = \frac{\min\limits_{1 \leq i \leq k} K_i}{m}.$$

Шаг 2. Вводим общую нумерацию клеток (последовательно от одного слоя к другому) и с каждой клеткой связываем набор попавших в неё векторов из X.

Шаг 3. Формируем таблицу «весов» векторов множества X. Здесь под «весом» вектора x понимаем число вхождений x в множество X.

Шаг 4. Формируем множество начальных (стартовых) векторов S для запуска процедуры «среднего сдвига». Для каждой клетки, которая содержит векторы из X, вычисляем вектор средних значений по всем точкам, попавшим в эту клетку. Совокупность полученных таким образом векторов образует множество S.

Шаг 5. Для каждого вектора $s \in S$ находим моду s^* оценки плотности распределения $\hat{f}_N(x)$, связную с s. Из найденных мод формируется множество $Z_0 = \{s^* | s \in S,\ \hat{f}_N(s^*) \geq \varepsilon\}$. По мере нахождения мод, заполняем множество $\bar{S} = \bigcup_{s \in S}\{\bar{s}_0 = s, \dots, \bar{s}_i = \bar{s}_{i-1} + m_h(\bar{s}_{i-1}), \dots, s^*\}$, которое содержит все точки, в которых вычислялось значение $\hat{f}_N(x)$.

Шаг 6. Связываем каждую точку $x \in X$ с ближайшей точкой из множества $\bar{S}$, используя для уменьшения количества вычислений введённую сеточную структуру. В результате множество X разбивается на кластеры в соответствии с определением 2.5.

Шаг 7. Выделяем многомодовые кластеры в соответствии с определением 2.6. Если искомая траектория P, соединяющая $Q(x_1^*)$ и $Q(x_2^*)$, существует, она проходит через общую границу $Q(x_1^*)$ и $Q(x_2^*)$. Поэтому для её нахождения выбираем точки $x_1 \in Q(x_1^*)$ и $x_2 \in Q(x_2^*)$, расположенные близко к общей границе, и проверяем выполнение условий из определения 2.6 для $P = [x_1, \dots, x_1^*] \cup [x_1, x_2] \cup [x_2, \dots, x_2^*]$. Здесь $[x_1, x_2]$ – отрезок прямой, соединяющей точки x_1 и x_2. Для нахождения точки траектории с наименьшей плотностью последовательно проверяем точки отрезка $[x_1, x_2]$ с шагом h.

При программной реализации последнего этапа алгоритма использовано наблюдение, что искомая траектория P, соединяющая кластеры, не может проходить через ячейки, отнесённые к «шуму» или не содержащие точек исходной выборки. Поэтому для нахождения P выбираем пары соседних ячеек, отнесённых к разным компонентам связности. После этого для каждой выбранной пары ячеек

находим наиболее близкие точки $x_1 \in Q(x_1^*)$ и $x_2 \in Q(x_2^*)$, и проверяем выполнения условий определения 2.6 для отрезка прямой, соединяющей x_1 и x_2. Блок-схема последнего шага предложенногго алгоритма (выделения многомодовых кластеров) приведена на рисунке 2.2.

2.5. Исследование алгоритма методом статистического моделирования

Разработанный алгоритм MeanSC исследован на модельных данных и реальных изображениях. Многочисленные эксперименты показали, что алгоритм вычислительно эффективен и способен выделять кластеры сложной структуры (формы, размера, плотности).

Все алгоритмы, разработанные в ходе выполнения диссертационной работы, реализованы на языке программирования C++ в среде Microsoft Visual Studio 2019 и включены в пакет программ «Image ProcessingToolkit», подробно описанный в разделе 5.3. Для эффективного использования многоядерных процессоров использован стандарт OpenMP. Обработка осуществлялась на ПЭВМ с процессором Intel Core i3-8100 (4 ядра по 3.6 ГГц) и 16 Гбайт оперативной памяти.

Для оценки качества результатов кластеризации введём следующее определение.

Определение 2.7. *Пусть для набора данных $X = \{x_1, \dots, x_N\}$ известно эталонное разбиение $\boldsymbol{g}^*: X \to \{G_0^*, \dots, G_M^*\}$. Тогда для произвольного разбиения $\boldsymbol{g}: X \to \{G_0, \dots, G_K\}$ установим соответствие $\gamma(G): \{G_0^*, \dots, G_M^*\} \to \{G_0, \dots, G_K, \emptyset\}$, при котором выполняется $\forall (G \neq \bar{G}): \gamma(G) = \gamma(\bar{G}) \Leftrightarrow \gamma(G) = \emptyset$ и достигается наибольшее значение выражения*

$$n_\gamma = \sum_{i=1}^{M} [|C_i^* \cap \gamma(C_i^*)| \cdot I(\gamma(C_i^*) \neq \emptyset)],$$

где $I(\cdot)$ – характеристическая функция. Тогда точность кластеризации (acc) определяется по следующей формуле:

$$acc = \frac{n_\gamma}{N} * 100\%.$$

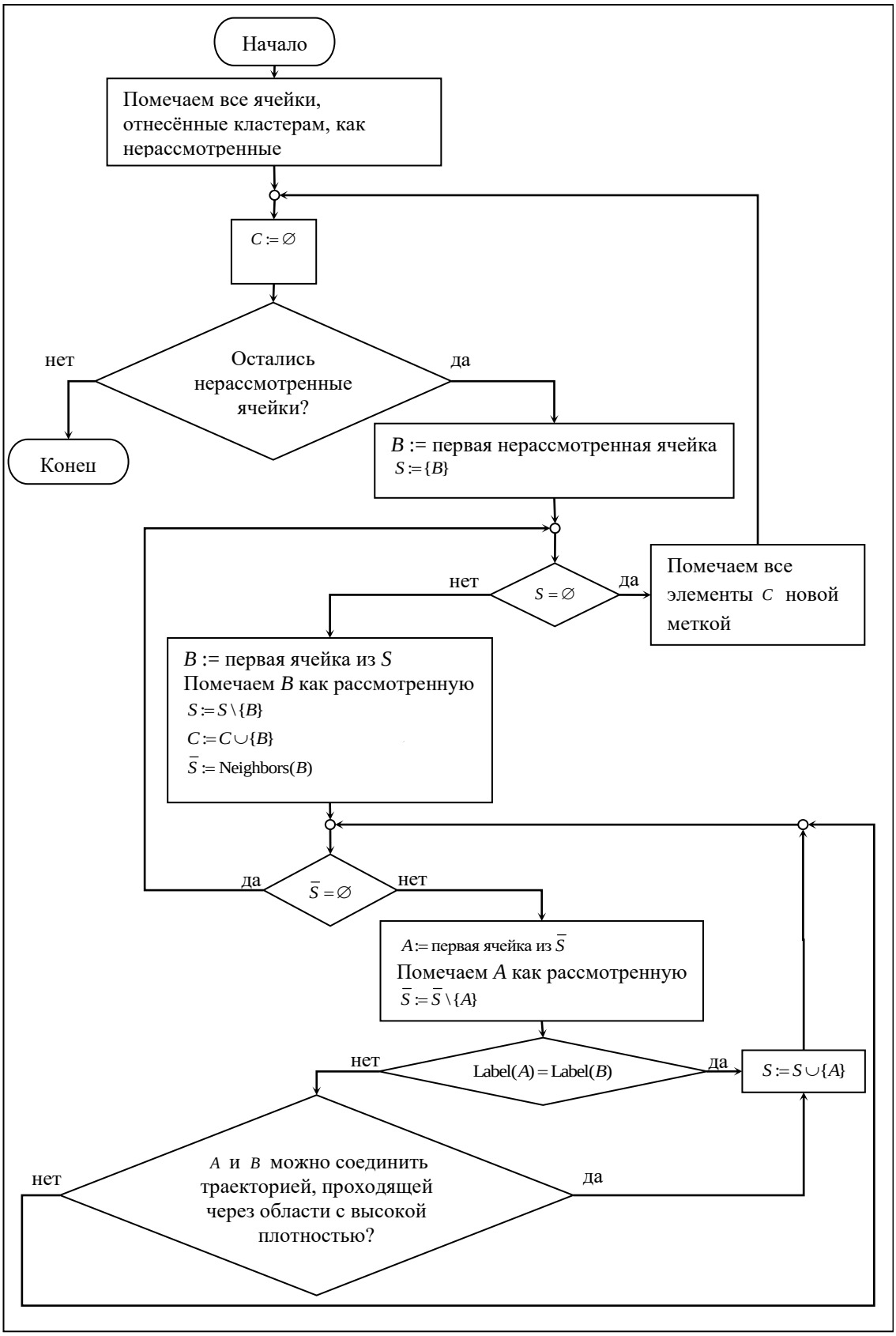


Рисунок 2.2 – Блок-схема последнего шага алгоритма MeanSC.

Обозначения: A, B – элементы сеточной структуры (ячейки); $\text{Label}(B)$ – текущая метка ячейки B; $\text{Neighbors}(B)$ – список нерассмотренных ячеек, соседних с B; C – формируемый многомодовый кластер; S, $\bar{S}$ – вспомогательные списки ячеек

При исследовании алгоритма MeanSC методом статистического моделирования использовалось три модельных набора данных. Для удобства интерпретации результатов генерировались только двумерных данные в интервале $[0,255]$ по каждой размерности.

Модельный набор данных 1 состоит из 1400 точек, сгруппированных в два класса. Первый класс содержит 400 точек и описывается нормальным распределением с вектором математического ожидания $\mu = (125,125)$ и ковариационной матрицей $\Sigma = \begin{pmatrix} 21^2 & 0 \\ 0 & 21^2 \end{pmatrix}$. Второй класс включает 1000 точек, равномерно распределённых по кольцу с центром в точке $(125,125)$ и радиусами $r = 80$, $R = 125$. Сложность модели заключается в том, что классы линейно неразделимы и имеют одинаковые вектора математического ожидания.

Модельный набор данных 2 состоит из трёх классов, включающих 300, 400 и 400 точек соответственно. Первый класс описывается нормальным распределением с вектором математического ожидания $\mu = (125,125)$ и ковариационной матрицей $\Sigma = \begin{pmatrix} 14^2 & 0 \\ 0 & 14^2 \end{pmatrix}$. Точки второго класса равномерно распределены в области $R_1 \backslash R_2$, где R_1 – круг с центром в точке $(115,140)$ и радиусом 100, а R_2 – круг с центром в точке $(115,190)$ и радиусом 100. Третий класс включает точки, равномерно распределённые в области $R_1 \backslash R_2$, где R_1 – круг с центром в точке $(135,110)$ и радиусом 100, а R_2 – круг с центром в точке $(85,190)$ и радиусом 100.

Модельный набор данных 3. Набор состоит из двух равновероятных классов, имеющих форму спиралей. Каждый класс состоит из 100 точек. Сложность модели заключается в том, что расстояние между точками спирали увеличивается по мере отдаления от центра, при этом расстояние между спиралями остаётся неизменным.

Модельные наборы данных 1-3 и их эталонные разбиения представлены на рисунке 2.3. Точность кластеризации этих наборов данных алгоритмом MeanSC составила 100% при использовании следующих параметров:

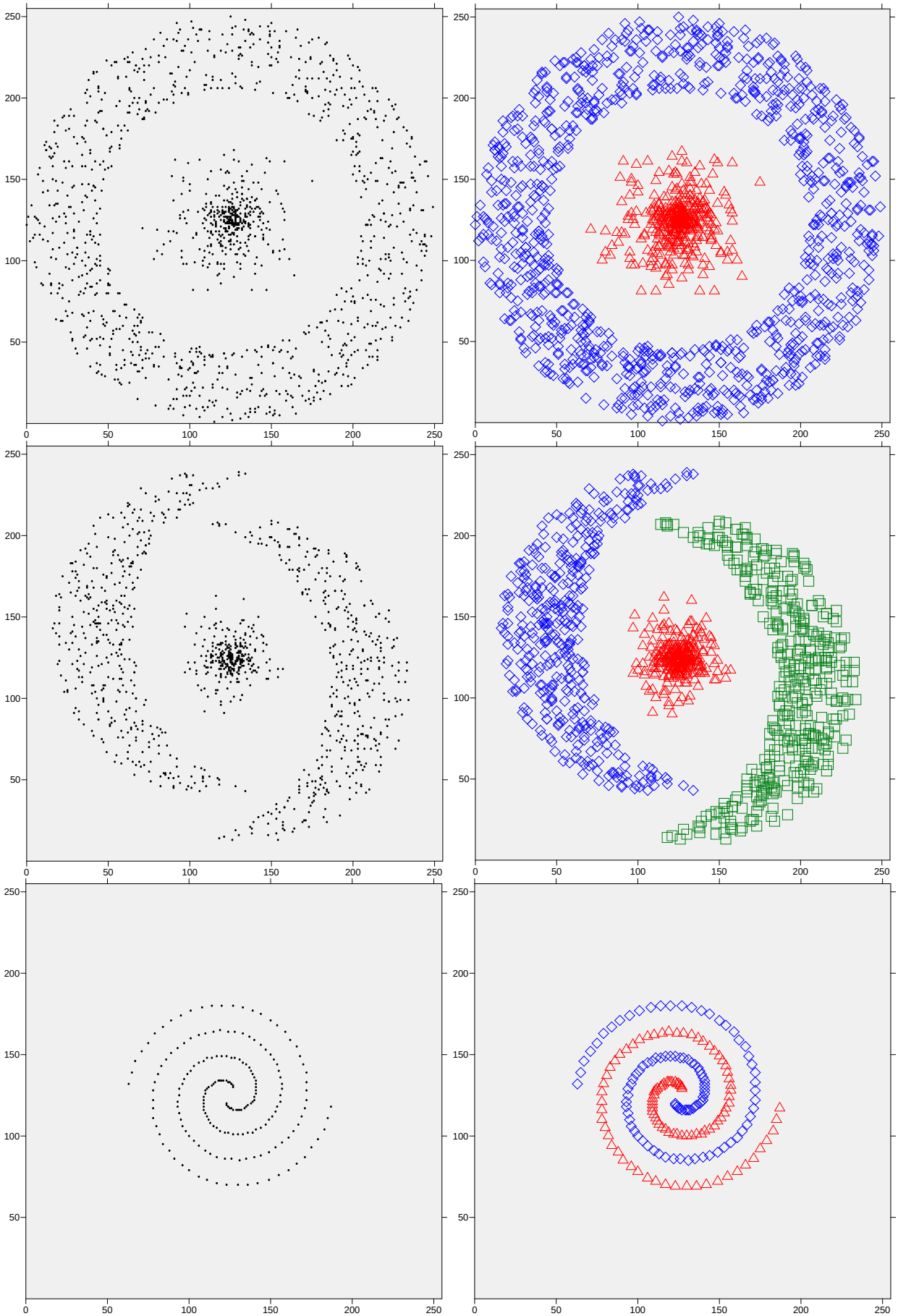


Рисунок 2.3 – Модельные наборы данных 1-3 (слева) и их эталонное разбиение (справа)

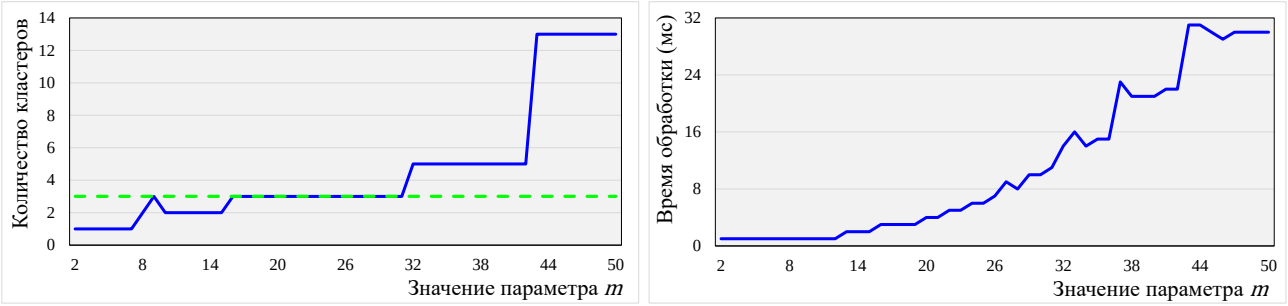
$m \in \{4,7,10\}$, $\varepsilon = 0$, $T = 0.5$ для набора 1; $16 \leq m \leq 19$, $\varepsilon = 0$, $T = 0.8$ для набора 2 и $22 \leq m \leq 31$, $\varepsilon = 0$, $T = 0.4$ для набора 3. Это доказывает применимость разработанного алгоритма для разделения линейно неразделимых классов сложной структуры.

Алгоритм MeanSC зависит от трёх параметров. Параметр ε регулирует порог отсечения «шума» и даёт возможность отбросить небольшие кластеры, тем самым повысив точность выделения кластеров большого размера. Это полезно при обработке изображений, содержащих мелкие объекты (например, технику на полях или отдельные небольшие строения).

Параметр m напрямую влияет на значение параметра сглаживания h оценки плотности $\hat{f}_N(x)$. С ростом m значение h уменьшается, что приводит к росту числа кластеров. Кроме того, параметр m определяет количество разбиений пространства признаков по каждой из осей, что оказывает значительное влияние на время обработки. На рисунке 2.4 представлены графики зависимости числа кластеров и времени обработки (в мс) алгоритмом MeanSC от значения параметра m для модельного набора данных 2 при фиксированных значениях параметров $\varepsilon = 0$ и $T = 0.8$. Зелёная штриховая линия обозначает число классов, присутствующих в данных.

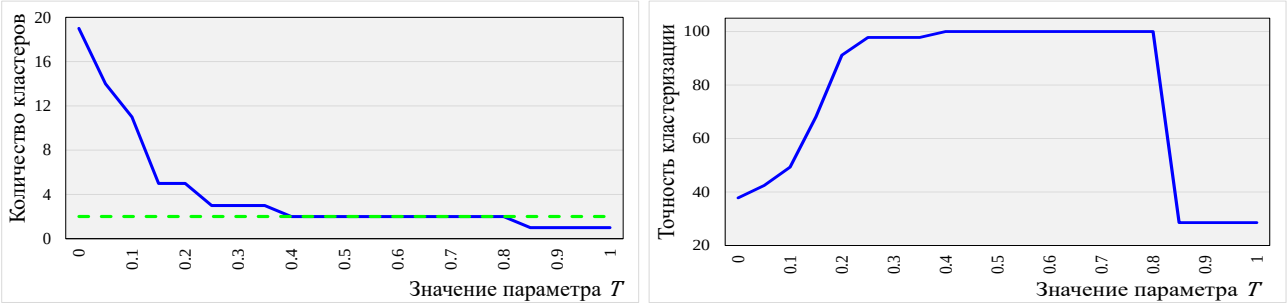
Параметр T используется для объединения компонент связности на последнем шаге алгоритма. Он влияет на степень детальности результата. Графики зависимости числа кластеров и точности кластеризации алгоритмом MeanSC от параметра T для модельного набора данных 1 представлены на рисунке 2.5 Зелёная штриховая линия обозначает число классов, присутствующих в данных. Значения параметров $m = 7$ и $\varepsilon = 0$ были зафиксированы.

На рисунке 2.6 представлен результат обработки цветного изображения. Сегментация выполнялась в цветовом пространстве $R \times G \times B$. Каждый кластер соответствует однородной области на изображении. Цветовое пространство содержало 1283800 точек, средний «вес» точек равнялся 19.69. Обработка с параметрами $m = 13$, $\varepsilon = 0$, $T = 0.05$ заняла 0.178 с и позволила выделить 12 кластеров.



а *б*

Рисунок 2.4 – Зависимость количества кластеров (*а*) и времени обработки (в мс) (*б*) модельного набора данных 2 от значения параметра m

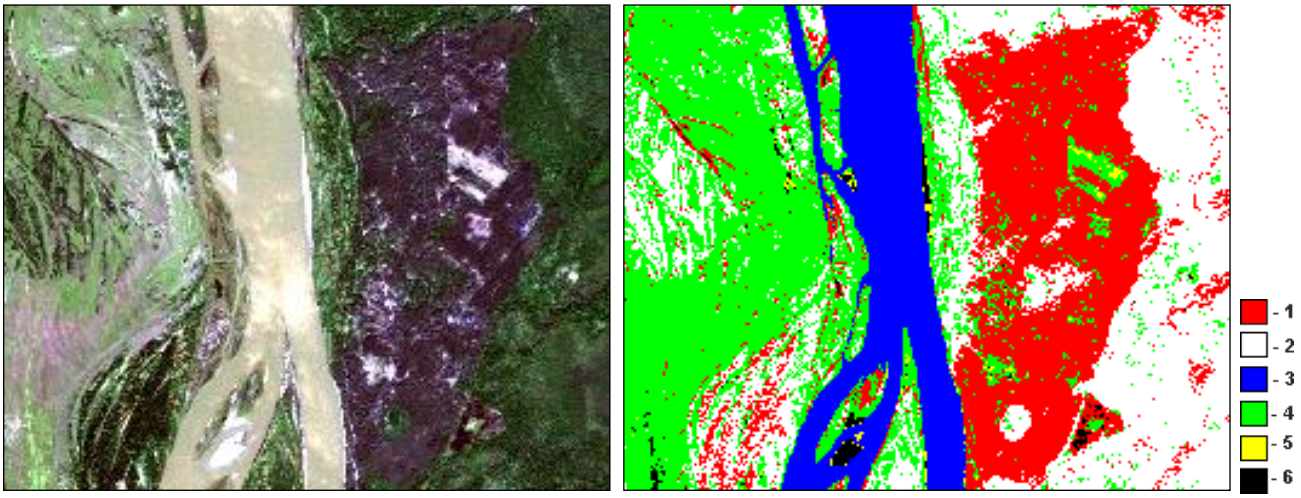


а *б*

Рисунок 2.5 – Зависимость количества кластеров (*а*) и точности кластеризации (*б*) модельного набора данных 1 от значения параметра T

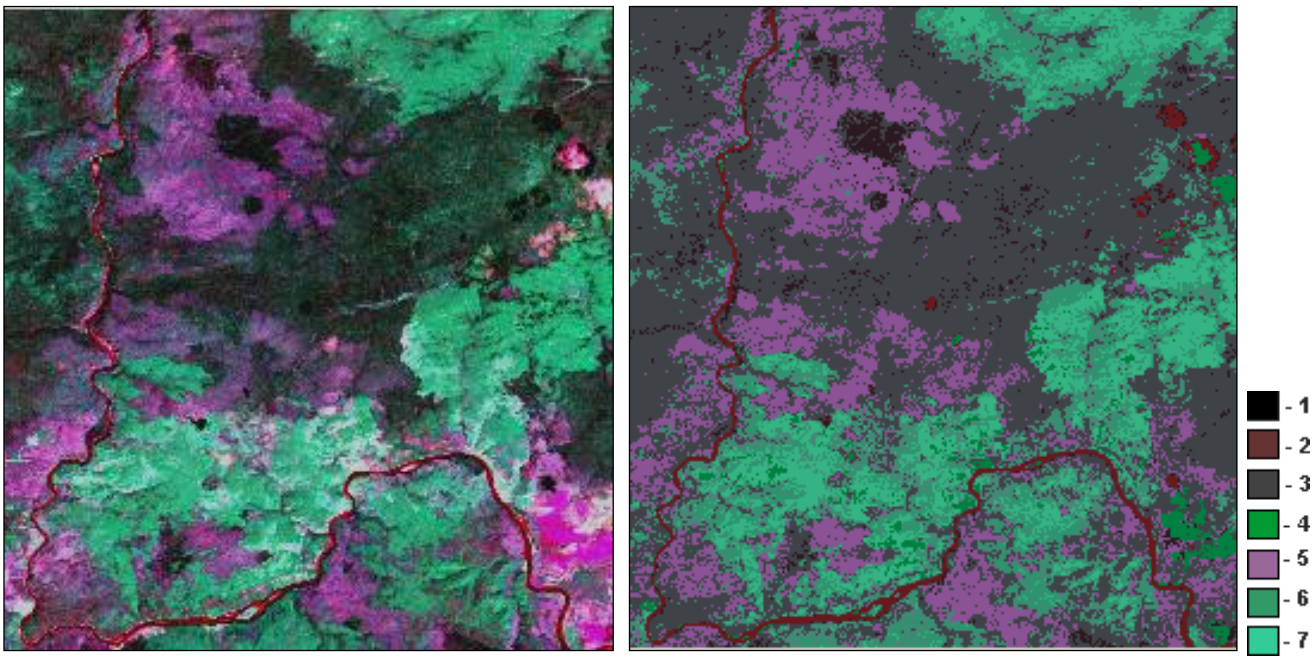
а *б*

Рисунок 2.6 – Изображение (*а*) и результат его сегментации алгоритмом MeanSC (*б*)



а *б*

Рисунок 2.7 – RGB-композит исходного фрагмента снимка ALOS (*а*) и результат его обработки (*б*). 1 – берёзовые насаждения; 2 – сосновые насаждения; 3 – поверхность воды; 4 – кустарники; 5 – участки, покрытые травой; 6 – песчаные отложения



а *б*

Рисунок 2.8 – RGB-композит исходного фрагмента снимка Landsat-7 (*а*) и результат его сегментации (*б*). 1 – перестойные (очень старые) берёзовые насаждения (возможно, вместе с тёмнохвойными), 2 – водная поверхность, 3 – тёмнохвойные насаждения (пихтовые и еловые), 4 – другие нарушенные территории, 5 – погибшие от шелкопряда насаждения, 6 – спелые берёзовые насаждения, 7 – молодые берёзовые насаждения

На рисунке 2.7 представлен результат обработки изображения, полученного со спутника ALOS. Для обработки выбран фрагмент размером 550×500 пикселей в 1-м, 2-м, 3-м и 4-м каналах. Исходное число элементов – 50000. Обработка алгоритмом MeanSC с параметрами $m = 17$, $\varepsilon = 0$, $T = 0.05$ заняла 0.2 с и позволила выделить шесть кластеров.

На рисунке 2.8 представлен фрагмент снимка повреждённых сибирским шелкопрядом темнохвойных лесов южной тайги Нижнего Приангарья, полученный со спутника LandSat-7. Обработке подвергался фрагмент размером 1001×1045 в 3-м, 4-м и 5-м каналах. Объём исходной выборки – 1046045. Обработка с параметрами $m = 20$, $\varepsilon = 0$, $T = 0.1$ заняла 0.5 с и позволила выделить семь кластеров.

Многочисленные экспериментальные исследования, проведённые как с модельными данными, так и с изображениями и цветными фотографиями, показывают, что предложенный алгоритм кластеризации MeanSC позволяет выделять кластеры разной структуры (формы, размера, плотности).

Выводы по главе

1. Приведена формальная постановка задачи кластеризации в рамках вероятностно-статистического подхода. Описан подход к её решению, основанный на процедуре «среднего сдвига». Приведён обзор способов выбора оптимального значения параметра сглаживания для оценки Розенблатта – Парзена.

2. Разработан и исследован вычислительно эффективный алгоритм кластеризации MeanSC на основе непараметрических оценок плотности Розенблатта – Парзена для сегментации мультиспектральных спутниковых изображений. Эффективность достигается за счёт введения сеточной структуры в пространстве признаков и переходу к рабочей выборке значительно меньшего объёма, в которой гарантированно содержатся представители всех классов, присутствующих на изображении. В отличие от предложенного в [115] алгоритма, при оценке плотности используется вся исходная выборка, что позволяет избежать неточностей при классификации. Предложенные оптимизации позволяют уменьшить время обработки изображений на два, а в некоторых случаях даже на три порядка.

3. Предложена оригинальная процедура распределения точек исходной выборки по кластерам, сформированным по точкам рабочей выборки. При распределении учитываются все точки, в которых вычислялась плотность при выполнении алгоритма. Это позволяет повысить точность выделения границ кластеров.

4. Предложенный алгоритм теоретически обоснован и исследован методом статистического моделирования на модельных данных и реальных изображениях, показана его эффективность.

ГЛАВА 3. АНСАМБЛЕВЫЙ АЛГОРИТМ КЛАСТЕРИЗАЦИИ EMEANSC (ENSEMBLE OF MEAN SHIFT CLASSIFIERS)

3.1. Ансамблевый подход к задаче автоматической классификации

Результаты выполнения алгоритмов автоматической классификации зачастую сильно зависят от значений настраиваемых параметров. Например, в алгоритме k-средних выбор начальных центров во многом определяет качество результата. Кроме того, структура и взаимное расположение кластеров, присутствующих на разных участках изображения, нередко различаются, поэтому для построения адекватного их разбиения требуются различные значения параметров алгоритма.

На практике, подбор вектора параметров, позволяющего получить хорошее разбиение всего изображения, требует значительных временных затрат и непрерывное участие эксперта. Для ускорения этого процесса всё чаще применяются ансамбли алгоритмов [133-139]. При этом используются как разбиения, полученные различными алгоритмами, так и результаты выполнения одного алгоритма (с различными значениями параметров, по различным подсистемам переменных и т.д.). Затем на основе построенного набора разбиений находится итоговое коллективное решение.

Идея объединения результатов выполнения нескольких простых алгоритмов для принятия сложных группировочных решений широко применяется в задачах распознавания образов [140], извлечения знаний из данных [141] и прогнозирования [142]. Она заключается в следующем.

Пусть с помощью набора группировочных функций $\mathbf{g} = \{g^{(1)}, \dots, g^{(L)}\}$ получено L частных решений $\mathbf{G} = \{G^{(1)}, \dots, G^{(L)}\} \subset \mathbb{G}$, где $\mathbb{G}$ – множество всех возможных разбиений X на произвольное число кластеров. Тогда итоговое коллективное решение (согласующая функция) представляет собой отображение $m: \mathbb{G}^L \to \mathbb{G}$.

Выбор наилучшей согласующей функции для объединения нескольких разбиений является сложной задачей. Простые схемы (например, голосование по большинству) в данном случае неприменимы, т.к. обычно нет однозначного соответ-

ствия между кластерами из разных разбиений. Известно несколько способов выбора наилучшей согласующей функции; наиболее распространёнными из них являются максимизация количества взаимной информации, которую разделяет итоговое коллективное решение с исходными частными решениями [138], и построение согласованной матрицы различий (или подобия) объектов [139]. Кроме того, существуют подходы, заключающиеся в сведении задачи выбора согласующей функции к более изученным задачам, таким как разбиение гиперграфа или кластеризация [136].

Рассмотрим эти подходы более подробно. Сопоставим каждой решающей функции g случайную переменную Y, принимающую значения из множества $\{1, \dots, M\}$ (где M – количество кластеров в разбиении G). Тогда нормализованное количество взаимной информации, которую разделяют группировочные решения $g^{(a)}$ и $g^{(b)}$, определяется по формуле [138]

$$NMI\left(Y^{(a)}, Y^{(b)}\right) = \frac{I(Y^{(a)}, Y^{(b)})}{\sqrt{H(Y^{(a)}, Y^{(b)})}},$$

где $I(Y^{(a)}, Y^{(b)})$ – количество взаимной информации между $Y^{(a)}$ и $Y^{(b)}$; $H(\cdot)$ – энтропия. Выборочной оценкой NMI служит величина

$$\phi^{(NMI)}\left(g^{(a)}, g^{(b)}\right) = \frac{\sum_{i,j} n_{i,j}^{(a),(b)} \log\left(\frac{N\, n_{i,j}^{(a),(b)}}{n_i^{(a)} n_j^{(b)}}\right)}{\sqrt{\left(\sum_i n_i^{(a)} \log\frac{n_i^{(a)}}{N}\right)\left(\sum_j n_j^{(b)} \log\frac{n_j^{(b)}}{N}\right)}},$$

где $i \in \{1, \dots, M^{(a)}\}$, $j \in \{1, \dots, M^{(b)}\}$; $n_i^{(a)}$ – число объектов в кластере $C_i^{(a)} \in G^{(a)}$; $n_j^{(b)}$ – число объектов в кластере $C_j^{(b)} \in G^{(b)}$; $n_{i,j}^{(a),(b)}$ – число объектов в пересечении кластеров $C_i^{(a)}$ и $C_j^{(b)}$.

Количество взаимной информации, которую разделяют согласующая функция m и набор решающих функций $\mathbf{g} = \{g^{(1)}, \dots, g^{(L)}\}$ находится по формуле

$$\phi^{(ANMI)}(m, \mathbf{g}) = \frac{1}{L}\sum_i \phi^{(NMI)}(m, g^{(i)}).$$

Тогда оптимальная согласующая функция m^{opt} равна

$$m^{\text{opt}} = \arg\max_{m} \phi^{(ANMI)}(m, \mathbf{g}).$$

В соответствии с другим способом [139], для нахождения согласующей функции строится согласованная матрица различий объектов. Для этого по каждому разбиению $G^{(a)} \in \mathbb{G}$, полученному при помощи решающей функции $g^{(a)}$, строится соответствующая матрица различий объектов $H(G^{(a)})$ размером $N \times N$ следующим образом:

$$H^{(i,j)}(G^{(a)}) = \begin{cases} 0, & \text{если } x^{(i)} \text{ и } x^{(j)} \text{ принадлежат одному кластеру в } G^{(a)}, \\ 1, & \text{иначе.} \end{cases}$$

Из матриц, построенных для набора разбиений $\mathbf{G}$, формируется согласованная матрица различий

$$H = \frac{1}{L}\sum_{i=1}^{L} H^{(i)}.$$

Элементы матрицы H характеризуют вероятность отнесения пар объектов из X в разные кластеры при использовании решающих функций из $\mathbf{g}$.

Для построения итогового группировочного решения, согласованная матрица различий используется как матрица расстояний между объектами для любого подходящего алгоритма кластеризации. Например, в [143] предлагается использовать эволюционный алгоритм группировки, а в [139] – агломеративный метод построения дендрограммы.

Существует несколько методов, сводящих задачу поиска согласующей функции к задаче разбиения гиперграфа [136, 138]. В [138] описаны два таких метода. В соответствии с первым из них, вершинам гиперграфа соответствуют элементы исходной выборки, а рёбрам – кластеры из $G^{(1)} \cup \ldots \cup G^{(L)}$. Тогда задача заключается в разбиении построенного гиперграфа по минимальному числу гиперрёбер. Таким образом, нахождение оптимальной согласующей функции сводится к задаче разбиения гиперграфа. В соответствии со вторым методом, вершинами разбиваемого гиперграфа являются кластеры из $G^{(1)} \cup \ldots \cup G^{(L)}$. Для вычисления весов рёбер (меры схожести кластеров) используется индекс Джаккарда (отношение меры

пересечения двух кластеров к мере их объединения). После декомпозиции элементы исходной выборки распределяются по метакластерам с помощью согласованной матрицы различий. Кроме того, в [136] описан метод построения и последующей декомпозиции двудольного графа, вершинами которого являются как элементы исходной выборки, так и найденные кластеры. Вес ребра (u, v) определяется следующим образом: если существует разбиение $G^{(a)}$, в котором объект u отнесён к кластеру v, то вес ребра (u, v) равен 1; во всех остальных случаях вес ребра (u, v) равен 0.

В [135] предлагается строить коллективное решение не по всем имеющимся вариантам группировки. Разбиения для формирования ансамбля выбираются при помощи «качества» и «разнообразия», которые вычисляются на основе меры количества взаимной информации.

Метод выбора согласующей функции на основе максимизации количества взаимной информации является вычислительно сложным. Методы на основе разбиения гиперграфа также характеризуются высокой трудоёмкостью. Поэтому в настоящей работе коллективное решение строится на основе согласованной матрицы различий.

3.2. Исследование свойств ансамбля, построенного с помощью согласованной матрицы различий

Для исследования свойств выбранного метода формирования коллективного решения рассмотрим его вероятностную модель [139].

Предположим, что имеется некоторая скрытая (непосредственно не наблюдаемая) переменная U, которая задает принадлежность каждого объекта $x \in X$ к некоторому из $M \geq 2$ классов. Каждый класс характеризуется определённым законом условного распределения $p(x|U = r) = f_r(x), r = 1, \dots, M$. Рассмотрим следующую вероятностную модель генерации данных. Пусть для каждого объекта определяется класс, к которому он относится, в соответствии с априорными вероятно-

стями $P_r = \mathbb{P}(U = r)$, $r = 1, \dots, M$, где $\sum_{i=1}^{L} P_r = 1$. Затем в соответствии с распределением $f_r(x)$ определяется значение x (процедура проводится независимо для каждого объекта).

Пусть с помощью некоторого алгоритма кластерного анализа μ строится разбиение множества объектов X на M подмножеств. Поскольку нумерация кластеров не играет роли, удобнее рассматривать отношение эквивалентности, т.е. указывать, относит ли алгоритм μ каждую пару объектов в один и тот же класс, либо в разные классы.

Выберем произвольную пару $x^{(i)}$ и $x^{(j)}$ различных объектов выборки. Определим для них величину

$$H^{(i,j)}(\mu) = \begin{cases} 0, & \text{если объекты } x^{(i)} \text{ и } x^{(j)} \text{ отнесены в один кластер,} \\ 1, & \text{иначе.} \end{cases}$$

Пусть $P_U = \mathbb{P}\big(U\big(x^{(i)}\big) \neq U(x^{(j)})\big)$ – вероятность отнесения объектов к различным классам. Например, при $M = 2$ указанная вероятность равна

$$\begin{aligned} P_U = 1 &- \mathbb{P}\big(U\big(x^{(i)}\big) = 1 | x^{(i)}\big)\mathbb{P}\big(U\big(x^{(j)}\big) = 1 | x^{(j)}\big) \\ &- \mathbb{P}\big(U\big(x^{(i)}\big) = 2 | x^{(i)}\big)\mathbb{P}\big(U\big(x^{(j)}\big) = 2 | x^{(j)}\big) = \\ &= 1 - \sum_{r=1}^{2} \frac{f_r\big(x^{(i)}\big) f_r\big(x^{(j)}\big) P_r^2}{p(x^{(i)}) p(x^{(j)})}, \end{aligned}$$

где $p(\omega) = \sum_{r=1}^{2} f_r(\omega) P_r$, $\omega = x^{(i)}, x^{(j)}$.

Обозначим вероятность ошибки, которую может совершить алгоритм μ при классификации $x^{(i)}$ и $x^{(j)}$, через $P_{\text{er}}(\mu)$, тогда

$$P_{\text{er}}(\mu) = \begin{cases} P_U, & \text{если } H^{(i,j)}(\mu) = 0, \\ 1 - P_U, & \text{если } H^{(i,j)}(\mu) = 1. \end{cases}$$

Легко заметить, что

$$P_{\text{er}}(\mu) = \Big(1 - H^{(i,j)}(\mu)\Big) P_U + H^{(i,j)}(\mu)(1 - P_U) = P_U + (1 - 2P_U) H^{(i,j)}(\mu).$$

Алгоритм μ зависит от случайного вектора параметров $\Theta \in \boldsymbol{\Theta}$: $\mu = \mu(\Theta)$. Чтобы подчеркнуть зависимость результатов работы от параметра Θ, в дальнейшем будем обозначать $H^{(i,j)}(\mu(\Theta)) = H^{(i,j)}(\Theta)$, $P_{er}\big(\mu(\Theta)\big) = P_{\text{er}}(\Theta)$.

Пусть в результате L-кратного применения алгоритма μ со случайно и независимо отобранными параметрами $\theta^{(1)},\dots,\theta^{(L)}$ получен набор решений $G^{(1)},\dots,G^{(L)}$, которым соответствуют матрицы $H\left(\theta^{(1)}\right),\dots,H\left(\Theta^{(L)}\right)$. Для определённости, будем считать, что L – нечётно. Коллективным (ансамблевым) решением по большинству голосов будем называть функцию

$$\mathbf{H}\left(H(\theta^{(1)}),\dots,H(\theta^{(L)})\right)=\begin{cases}0, \text{ если } \dfrac{1}{L}\displaystyle\sum_{l=1}^{L}H(\theta^{(l)})<\dfrac{1}{2},\\ 1, \text{ иначе.}\end{cases}$$

В рамках описанной модели для выбранного способа построения коллективного решения справедливы следующие утверждения, доказательства которых приведены в [139].

Утверждение 3.1. *Математическое ожидание и дисперсия величины вероятности ошибки для алгоритма* $\mu(\Theta)$ *равны соответственно:*

$$\mathbb{E}_{\Theta}P_{\text{er}}(\Theta)=P_U+(1-2P_U)P_H,\quad \text{Var}_{\Theta}P_{\text{er}}(\Theta)=(1-2P_U)^2P_H(1-P_H),$$

где $P_H=\mathbb{P}(H(\Theta)=1)$.

Обозначим через $P_{\text{er}}(\Theta^{(1)},\dots,\Theta^{(L)})$ случайную функцию, принимающую при фиксированных аргументах значение, равное вероятности ошибки при классификации $x^{(i)}$ и $x^{(j)}$ с помощью ансамблевого алгоритма. Здесь через $\Theta^{(1)},\dots,\Theta^{(L)}$ обозначены статистические копии случайного вектора Θ. Рассмотрим поведение вероятности ошибки для коллективного решения.

Утверждение 3.2. *Математическое ожидание и дисперсия величины вероятности ошибки для коллективного решения равны соответственно:*

$$\mathbb{E}_{\Theta^{(1)},\dots,\Theta^{(L)}}P_{\text{er}}\left(\Theta^{(1)},\dots,\Theta^{(L)}\right)=P_U+(1-2P_U)P_{\mathbf{H},L},$$

$$\text{Var}_{\Theta^{(1)},\dots,\Theta^{(L)}}P_{\text{er}}\left(\Theta^{(1)},\dots,\Theta^{(L)}\right)=(1-2P_U)^2P_{\mathbf{H},L}(1-P_{\mathbf{H},L}),$$

где $P_{\mathbf{H},L}=\mathbb{P}\left(\dfrac{1}{L}\displaystyle\sum_{l=1}^{L}H(\theta^{(l)})\dfrac{1}{2}\right)=\displaystyle\sum_{l=\left\lfloor\frac{L}{2}\right\rfloor+1}^{L}C_L^lP_H^l(1-P_H)^{L-l}$, $[\cdot]$ *означает целую часть числа.*

Воспользуемся следующей априорной информацией об алгоритме кластерного анализа. Будем считать, что ожидаемая вероятность ошибочной классификации $\mathbb{E}_{\Theta}P_{\text{er}}(\Theta)<1/2$ (то есть ожидается, что алгоритм μ проводит классификацию с лучшим качеством, нежели алгоритм случайного равновероятного выбора). Из утверждения 1 следует, что выполняется один из двух вариантов: а) $P_H>1/2$ и $P_U>1/2$; б) $P_H<1/2$ и $P_U<1/2$. Рассмотрим, для определённости, первый случай.

Утверждение 3.3. *Если $\mathbb{E}_{\Theta}P_{\text{er}}(\Theta)<1/2$ и при этом $P_H>1/2$ и $P_U>1/2$, то при увеличении мощности ансамбля ожидаемая вероятность ошибочной классификации уменьшается, стремясь в пределе к $1-P_U$, а дисперсия величины вероятности ошибки стремится к нулю.*

Последнее утверждение позволяет сделать следующий вывод: при выполнении вполне естественных условий, использование ансамблевого подхода с решающей функцией, построенной на основе согласованной матрицы различий, позволяет повысить качество кластеризации.

3.3. Ансамблевый алгоритм кластеризации EMeanSC

Многочисленные эксперименты на модельных данных и реальных изображениях показали, что качество результатов кластеризации, получаемых с помощью алгоритма MeanSC, сильно зависит от значения параметра m. Поэтому для формирования ансамбля использовалось множество из L частных решений, полученных в результате выполнения алгоритма MeanSC с различными значениями параметра m. Выбранный способ построения коллективного решения может быть формализован следующим образом.

Пусть с помощью некоторого алгоритма кластеризации $\mu=\mu(\theta)$, зависящего от случайного вектора параметров $\theta\in\boldsymbol{\Theta}$ (где $\boldsymbol{\Theta}$ – некоторое допустимое множество параметров), получен набор частных решений $\mathbb{G}=\{G^{(1)},\dots,G^{(L)}\}$, где $G^{(l)}$ – l-й вариант кластеризации, соответствующий вектору параметров $\theta^{(l)}$ и содержащий $M^{(l)}$ кластеров.

Обозначим через $H(\theta^{(l)})$ бинарную матрицу $H(\theta^{(l)}) = \{H^{(i,j)}(\theta^{(l)})\}$ размерности $N \times N$, которая строится для l-го варианта разбиения следующим образом:

$$H^{(i,j)}(\theta^{(l)}) = \begin{cases} 0, & \text{если объекты } x^{(i)} \text{ и } x^{(j)} \text{ отнесены в один кластер,} \\ 1, & \text{иначе,} \end{cases}$$

где $i, j = 1, \dots, N$, $i \neq j$.

После построения L частных решений можно сформировать согласованную матрицу различий

$$\mathbf{H} = \{\mathbf{H}^{(i,j)}\}, \quad \mathbf{H}^{(i,j)} = \frac{1}{L}\sum_{l=1}^{L} H^{(i,j)}(\theta^{(l)}),$$

где $i, j = 1, \dots, N$. Величина $\mathbf{H}^{(i,j)}$ равна частоте отнесения $x^{(i)}$ и $x^{(j)}$ в разные кластеры в наборе группировок $\mathbb{G}$. Близкое к нулю значение величины означает, что данные объекты имеют большой шанс попадания в один и тот же кластер; близкое к единице значение говорит о том, что шанс оказаться в одном кластере у объектов незначителен.

После вычисления согласованной матрицы различий, для нахождения коллективного решения применяется стандартный агломеративный метод построения дендрограммы, который в качестве входной информации использует попарные расстояния между объектами [15]. При этом расстояние между кластерами объектов определяется по принципу «средней связи» (т.е. как среднее арифметическое попарных расстояний между объектами, отнесёнными к разным кластерам). Процесс объединения продолжается до тех пор, пока расстояние между ближайшими кластерами не превысит заданное пороговое значение.

В соответствии с описанной схемой на основе алгоритма MeanSC разработан ансамблевый алгоритм EMeanSC (Ensemble of MeanSC), который зависит от четырёх параметров: набора значений параметра сглаживания $\overline{m} = \{m^{(1)}, \dots, m^{(L)}\}$, минимального значения плотности ε, порога объединения кластеров T и порогового значения T_d. Алгоритм можно представить в виде следующей последовательности шагов.

Шаг 1. Для каждого значения параметра $m^{(l)} \in \overline{m}$ построить соответствующие разбиение, выполнив $\mathrm{MeanSC}(m^{(l)}, \varepsilon, T)$.

Шаг 2. На основе полученных разбиений построить согласованную матрицу различий **H**.

Шаг 3. Построить дендрограмму с помощью стандартного агломеративного метода, используя **H** в качестве матрицы расстояний. Для определения расстояния между кластерами использовать метод «средней связи». Построение продолжать до тех пор, пока расстояние между ближайшими группами не превысит пороговое значение T_d.

Стоит заметить, что при построении согласованной матрицы различий учитываются только метки, полученные на первом шаге. Если метки элементов $x^{(a)}$ и $x^{(b)}$ совпадают в каждом из разбиений, значение соответствующей ячейки $\mathbf{H}^{(a,b)}$ будет равно нулю и эти элементы однозначно попадают в один кластер. Поэтому, для экономии памяти и снижения вычислительных затрат, при нахождении итогового решения элементы, метки которых не отличаются ни в одном из разбиений, объединяются до начала формирования согласованной матрицы различий.

3.4. Исследование алгоритма методом статистического моделирования

При исследовании алгоритма EMeanSC методом статистического моделирования использовалось два модельных набора данных, а также распространённые данные по ирисам.

Модельный набор данных 4. Используется известная модель «бананы», построенная с помощью модуля PRTOOLS для Matlab. Набор включает два равновероятных класса. Значения признаков равномерно распределены по форме бананов и сдвинуты по каждой координате на случайную величину, характеризующуюся нормальным распределением с математическим ожиданием $\mu = 0$ и среднеквадратичным отклонением $\sigma = 0.7$. Для данного набора генерировалось по 200 точек для каждого класса.

Модельный набор данных 5. Модельные данные состоят из пяти классов, сильно пересекающихся в пространстве признаков. Первый класс содержит 2000 точек и описывается нормальным распределением с вектором математического

ожидания $\mu_1 = (125,125)$ и ковариационной матрицей $\Sigma_1 = \begin{pmatrix} 45^2 & 0 \\ 0 & 45^2 \end{pmatrix}$. Остальные классы содержат по 1000 точек и имеют нормальное распределение с векторами математического ожидания $\mu_2 = (150,180)$, $\mu_3 = (55,150)$, $\mu_4 = (100,55)$, $\mu_5 = (190,100)$ и одинаковыми ковариационными матрицами $\Sigma = \begin{pmatrix} 25^2 & 0 \\ 0 & 25^2 \end{pmatrix}$. Сложность модели заключается в значительном пересечении классов. Модельные наборы данных 4 и 5, а также их эталонные разбиения, представлены на рисунке 3.1.

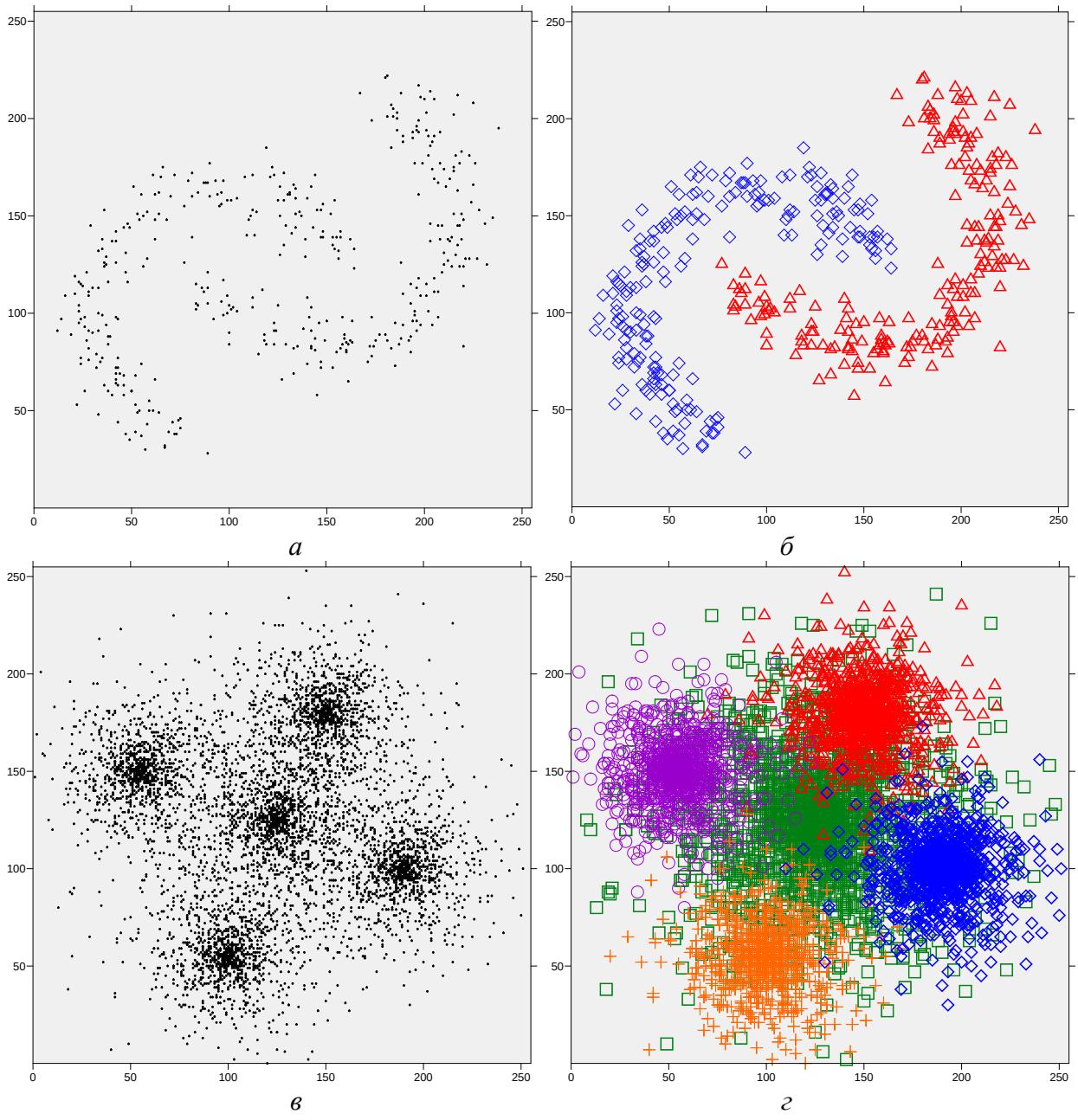


Рисунок 3.1 – Модельные наборы данных 4 и 5 (*а* и *в*) и их эталонные разбиения (*б* и *г*)

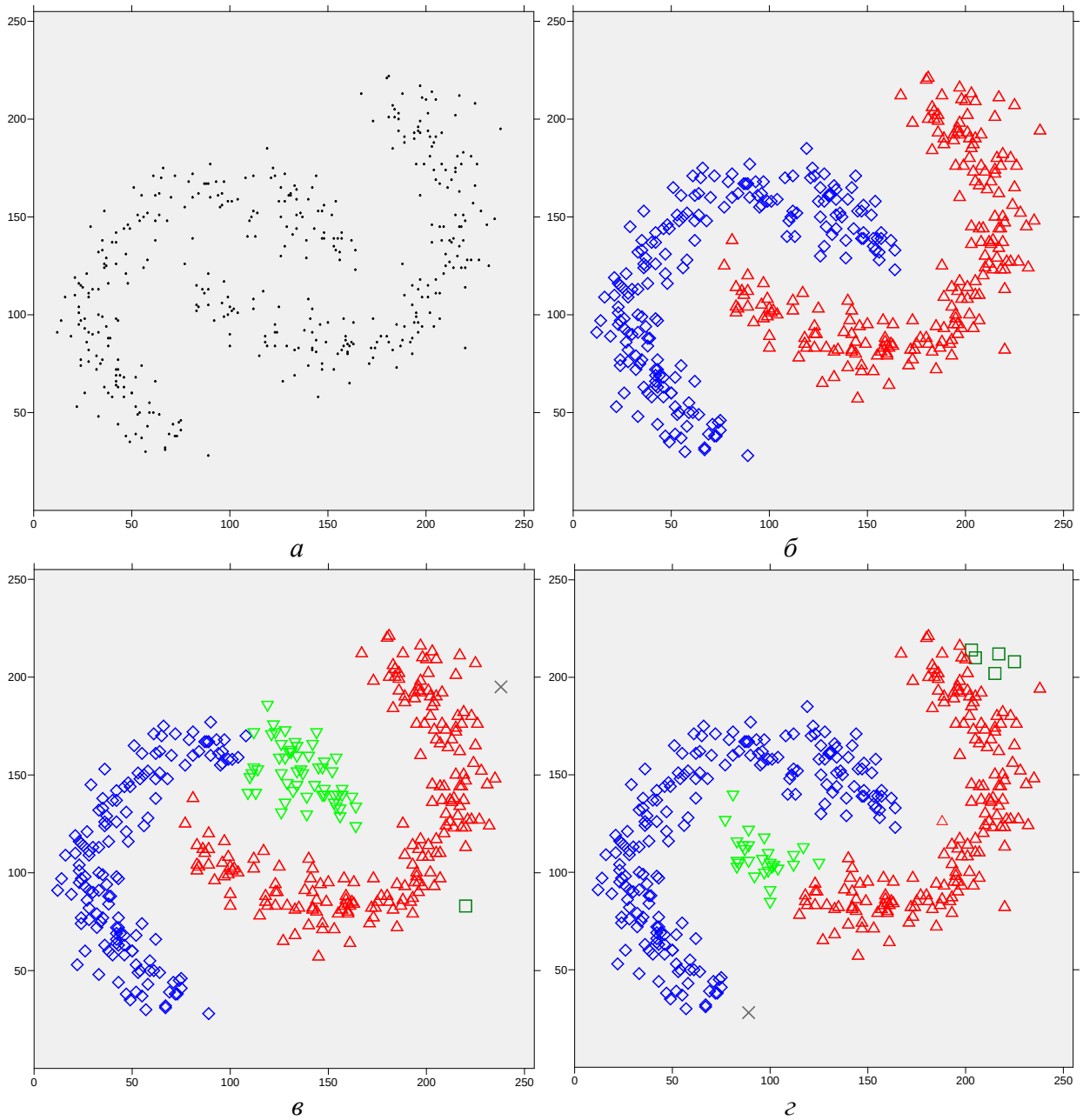
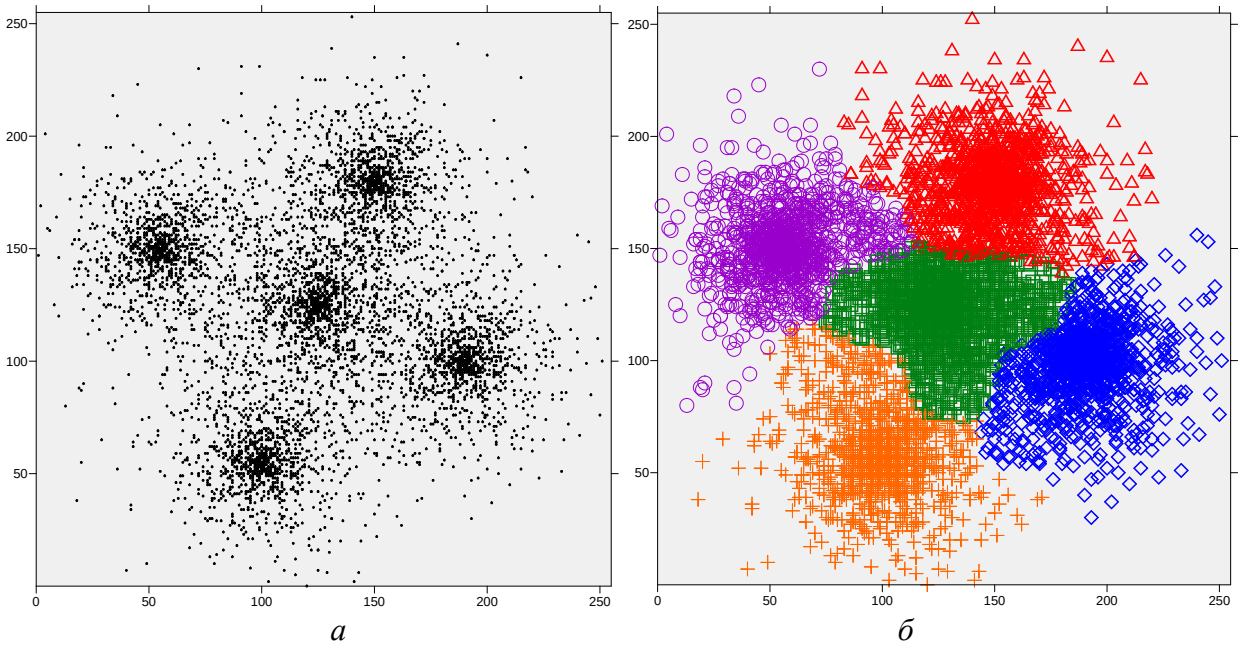


Рисунок 3.2 – Модельный набор данных 4 (*а*), результат выполнения ансамблевого алгоритма (*б*) и элементы ансамбля (*в* и *г*)

На рисунке 3.2 представлены результаты обработки модельного набора данных 4 алгоритмом EMeanSC с параметрами $\overline{m} = \{13, 14\}$, $\varepsilon = 0$, $T = 0.5$, $T_d = 0.6$ (ансамблевое решение, точность кластеризации 99.75%) и алгоритмом MeanSC с параметрами $m \in \{13, 14\}$, $\varepsilon = 0$, $T = 0.5$ (элементы ансамбля, точность кластеризации 85.5% и 91.25% соответственно). Несложно заметить, что все элементы ансамбля содержат грубые ошибки (искомые классы раздроблены на несколько кла-

стеров), а в ансамблевом решении, построенном на их основе, такие ошибки отсутствуют. Дробление классов связано с маленьким расстоянием между классами и низкой плотностью внутри них. Алгоритм EMeanSC позволяет успешно исправлять подобные ошибки. Эксперименты показали, что алгоритм MeanSC позволяет получить разбиение набора данных 4 с точностью кластеризации 99.75% только при одном наборе параметров ($m = 9$, $\varepsilon = 0$, $T = 0.6$).

На рисунке 3.3 представлен результат обработки модельного набора данных 5 алгоритмом EMeanSC с параметрами $\bar{m} = \{5,7\}$, $\varepsilon = 0$, $T = T_d = 0.4$. Время обработки составило 0.13 с, точность кластеризации – 86.7%. Видно, что алгоритм EMeanSC позволяет успешно разделять пересекающиеся классы.



а *б*

Рисунок 3.3 – Модельный набор данных 5 (*а*) и результат выполнения алгоритма EMeanSC (*б*)

Таблица 3.1 – Результаты обработки данных по ирисам

	$\lvert C_i^O \rvert$	Алгоритм EMeanSC				Алгоритм GCOD			
		$\lvert C_i^S \rvert$	$\lvert C_i^O \cap C_i^S \rvert$	Точность, %	Покрытие, %	$\lvert C_i^S \rvert$	$\lvert C_i^O \cap C_i^S \rvert$	Точность, %	Покрытие, %
$i=1$	50	50	50	100	100	50	50	100	100
$i=2$	50	46	46	100	92	19	18	94.7	36
$i=3$	50	54	50	92.6	100	81	49	60.5	98

Для сравнения разработанного алгоритма с алгоритмом GCOD использовались распространённые данные по ирисам [144, 145]. Данные состоят из 150 четырехмерных точек, сгруппированных в три класса по 50 точек. В соответствии с [106], обозначим: $|C_i^O|$– число точек, принадлежащих i -му классу, $|C_i^S|$ – число точек, отнесенных в соответствующий кластер, выделенный алгоритмом EMeanSC. Тогда $|C_i^O \cap C_i^S|/|C_i^S|$ и $|C_i^O \cap C_i^S|/|C_i^O|$ – точность алгоритма и покрытие соответственно. Результаты выполнения алгоритма EMeanSC с параметрами $\overline{m} = \{15,16\}$, $\varepsilon = 0$, $T = 0.15$, $T_d = 0.67$ и результаты, полученные в [106] с помощью алгоритма GCOD, представлены в таблице 3.1.

Выводы по главе

1. Приведена формальная постановка задачи формирования коллективного решения. Выполнен анализ известных стратегий выбора согласующей функции, которые позволяют объединять результаты кластеризации. Показано, что быстродействие, достаточное для обработки мультиспектральных спутниковых изображений, позволяет обеспечить только подход на основе согласованной матрицы различий

2. Предложен подход к построению ансамбля непараметрических алгоритмов кластеризации, основанных на оценках плотности Розенблатта – Парзена, с помощью согласованной матрицы различий. В рамках этого подхода на основе алгоритма MeanSC разработан ансамблевый алгоритм кластеризации EMeanSC, позволяющий осуществлять обработку мультиспектральных спутниковых изображений в диалоговом режиме.

3. Предложенный алгоритм исследован методом статистического моделирования. Показано, что он позволяет значительно упростить процедуру подбора параметров.

ГЛАВА 4. ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ ПРЕДЛОЖЕННЫХ АЛГОРИТМОВ

На данный момент в литературе предложено большое количество непараметрических алгоритмов кластеризации. Однако программная реализация большинства из них отсутствует. Как следствие, отсутствует возможность не только практического использования этих алгоритмов, но и сравнения их с алгоритмами, разработанными другими авторами.

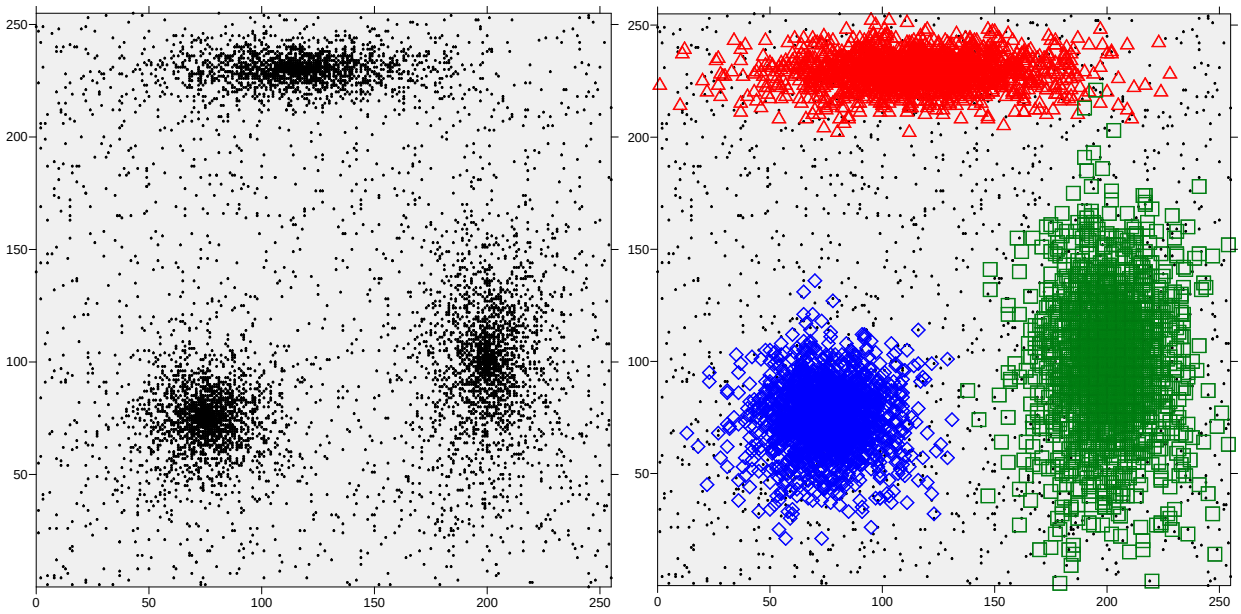
В ходе диссертационной работы было выполнено экспериментальное сравнение предложенных алгоритмов с алгоритмами из пакетов с открытым исходным кодом ELKI (https://elki-project.github.io) и Smile (https://haifengl.github.io/smile), предназначенных для анализа данных, а также с алгоритмами из коммерческого пакета для обработки спутниковых данных ITTVIS ENVI (http://www.harrisgeospatial.com/SoftwareTechnology/ENVI.aspx).

Все эксперименты, описанные в данной главе, выполнялись на ПЭВМ с процессором Intel Core i3-8100 (4 ядра по 3.6 ГГц) и 16 Гбайт оперативной памяти.

4.1. Экспериментальное исследование на модельных данных

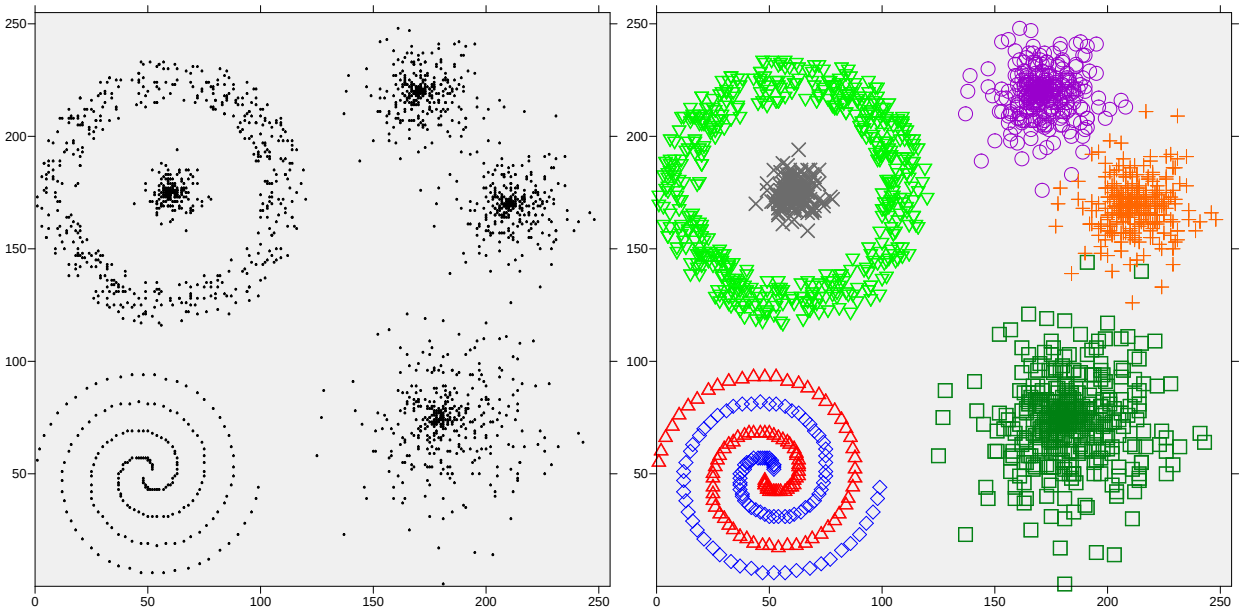
Для оценки качества кластеризации использовались семь модельных наборов данных. Наборы 1-5 описаны в параграфах 2.5 и 3.4.

Модельный набор данных 6 содержит три равновероятных класса и класс-шум. Первые три класса описываются нормальным распределением с векторами математического ожидания $\mu_1 = (75,75)$, $\mu_2 = (115,230)$, $\mu_3 = (200,100)$ и ковариационными матрицами $\Sigma_1 = \begin{pmatrix} 20^2 & 0 \\ 0 & 20^2 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 40^2 & 0 \\ 0 & 10^2 \end{pmatrix}$, $\Sigma_3 = \begin{pmatrix} 20^2 & 0 \\ 0 & 40^2 \end{pmatrix}$. Класс «шум» равномерно распределён по всему пространству признаков. Для модели генерировалось по 2000 точек первых трёх классов и 1500 точек «шума». Модельный набор данных 6 и его эталонное разбиение представлены на рисунке 4.1. Здесь и далее точки, отнесённые к «шуму», обозначены на рисунке чёрными точками.



а *б*

Рисунок 4.1 – Модельный набор данных 6 (*а*) и его эталонное разбиение (*б*)



а *б*

Рисунок 4.2 – Модельный набор данных 7 (*а*) и его эталонное разбиение (*б*)

Модельный набор данных 7 включает семь классов. Первые четыре класса описываются нормальным законом с векторами математического ожидания $\mu_1 = (180,75)$, $\mu_2 = (170,220)$, $\mu_3 = (210,170)$, $\mu_4 = (60,175)$ и ковариационными матрицами $\Sigma_1 = \begin{pmatrix} 25^2 & 0 \\ 0 & 25^2 \end{pmatrix}$, $\Sigma_2 = \Sigma_3 = \begin{pmatrix} 15^2 & 0 \\ 0 & 15^2 \end{pmatrix}$, $\Sigma_4 = \begin{pmatrix} 7^2 & 0 \\ 0 & 7^2 \end{pmatrix}$. Элементы пятого класса равномерно распределены по кольцу с центром в точке

$(60, 175)$ и радиусами $r = 40$, $R = 60$. Последние два класса имеют форму спиралей. Для первого-седьмого классов генерировалось 400, 300, 300, 200, 500, 100 и 100 точек соответственно. Данный набор данных является очень сложным для непараметрических методов кластеризации, поскольку он включает линейно неразделимые классы и близко расположенные классы, описываемые нормальным распределением с разной дисперсией. Сгенерированный набор данных и его эталонное разбиение представлены на рисунке 4.2.

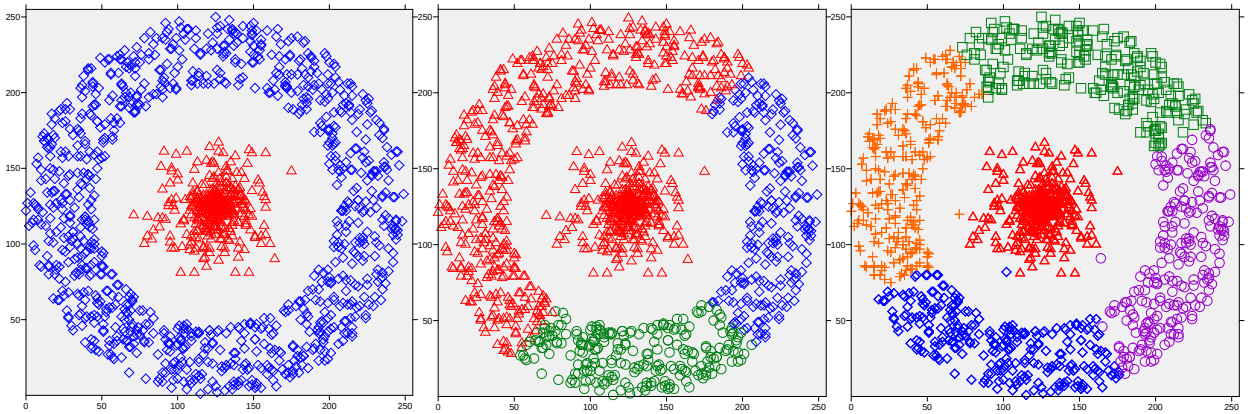
Представленные модельные наборы данных использовались для сравнения алгоритмов кластеризации MeanSC и EMeanSC, разработанных в ходе выполнения диссертационной работы, с известными непараметрическими алгоритмами DBSCAN, DENCLUE, k-средних и SLINK. Эти алгоритмы подробно описаны в первой главе диссертации, их реализации взяты из пакетов программ с открытым исходным кодом ELKI и SMILE. При настройке параметров алгоритмов преследовалась цель получения максимальной точности кластеризации (см. определение 2.7). Полученные значения точности кластеризации и время обработки представлены в таблице 4.1.

При обработке первых трех модельных наборов данных все алгоритмы, кроме DENCLUE и k-средних, удалось настроить для получения эталонного разбиения. Результаты обработки этих наборов данных представлены на рисунках 4.3-4.5. При обработке модельного набора данных 4 алгоритмами MeanSC и EMeanSC допущено по одной ошибке (рисунок 4.6, *б*). Немного менее точный результат продемонстрировали алгоритмы DBSCAN и SLINK (рисунок 4.6, *в* и *г*). Здесь и далее чёрными точками на результате выполнения SLINK отмечены кластеры размером менее 30 точек. Модельный набор данных 5 вызвал трудности у алгоритмов DBSCAN и SLINK, остальные алгоритмы позволили получить точность кластеризации порядка 86% (рисунок 4.7). Ошибки при разбиении этого набора данных связаны со значительным пересечением классов модели. Модельный набор данных 6 содержит класс-шум, успешно выделенный только с помощью алгоритмов MeanSC, EMeanSC и DBSCAN (рисунок 4.8). Остальные алгоритмы не позволили получить точность выше 80% (рисунки 4.9 и 4.10). При обработке модельного

набора данных 7 только алгоритмы MeanSC и EMeanSC позволили корректно выделить все классы модели (рисунок 4.11, *б* и *в*). Алгоритмы DBSCAN (рисунок 4.11, *г*), DENCLUE (рисунок 4.12, *б*) и k-средних (рисунок 4.12, *в*) позволили достаточно точно выделить только нормально распределённые классы. Алгоритм SLINK (рисунок 4.12, *г*), помимо классов с нормальным распределением, позволил выделить класс-кольцо.

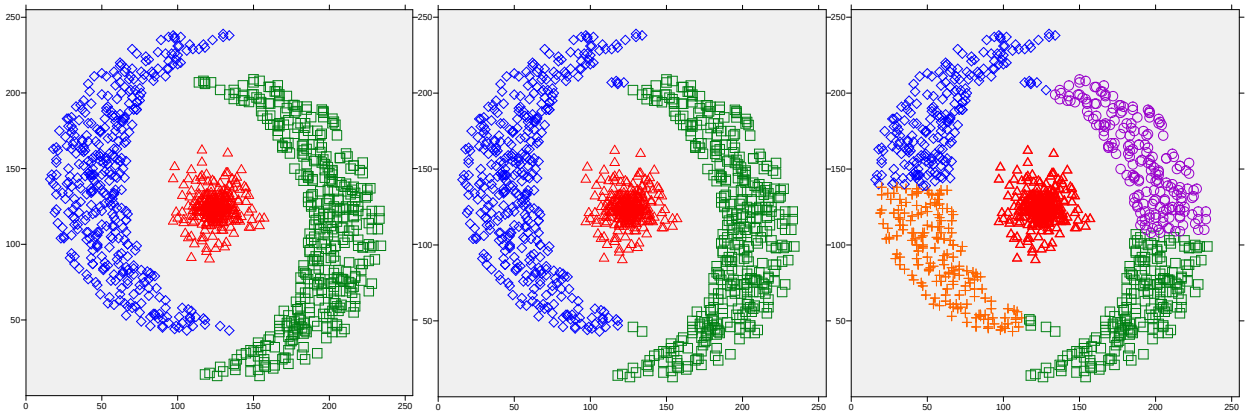
Таблица 4.1 – Точность кластеризации и время обработки модельных наборов данных

Алгоритм	Модельный набор данных						
	1	2	3	4	5	6	7
MeanSC	100% 0.002 с	100% 0.005 с	100% 0.004 с	99.75% 0.002 с	86.7% 0.01 с	89.18% 0.014 с	98.6% 0.017 с
EMeanSC	100% 0.018 с	100% 0.017 с	100% 0.007 с	99.75% 0.007 с	86.7 0.13 с	89.16% 0.19 с	98.7% 0.17 с
DBSCAN	100% 0.031 с	100% 0.016 с	100% 0.016 с	99% 0.016 с	65.2% 0.374 с	84.56% 1.17 с	90.63% 0.031 с
DENCLUE	65.14% 0.178 с	99.36% 0.09 с	50% 0.06 с	50.5% 0.14 с	85.9% 0.184 с	79.92% 3.26 с	83.8% 0.721 с
k-средних	47.79% 0.016 с	64.73% 0.015 с	53% 0.015 с	47.75% 0.016 с	84.67% 0.016 с	79.92% 0.085 с	78.99% 0.017 с
SLINK	100% 0.031 с	100% 0.016 с	100% 0.016 с	99% 0.016 с	54.08% 0.266 с	78.92% 0.415 с	91.08% 0.031 с



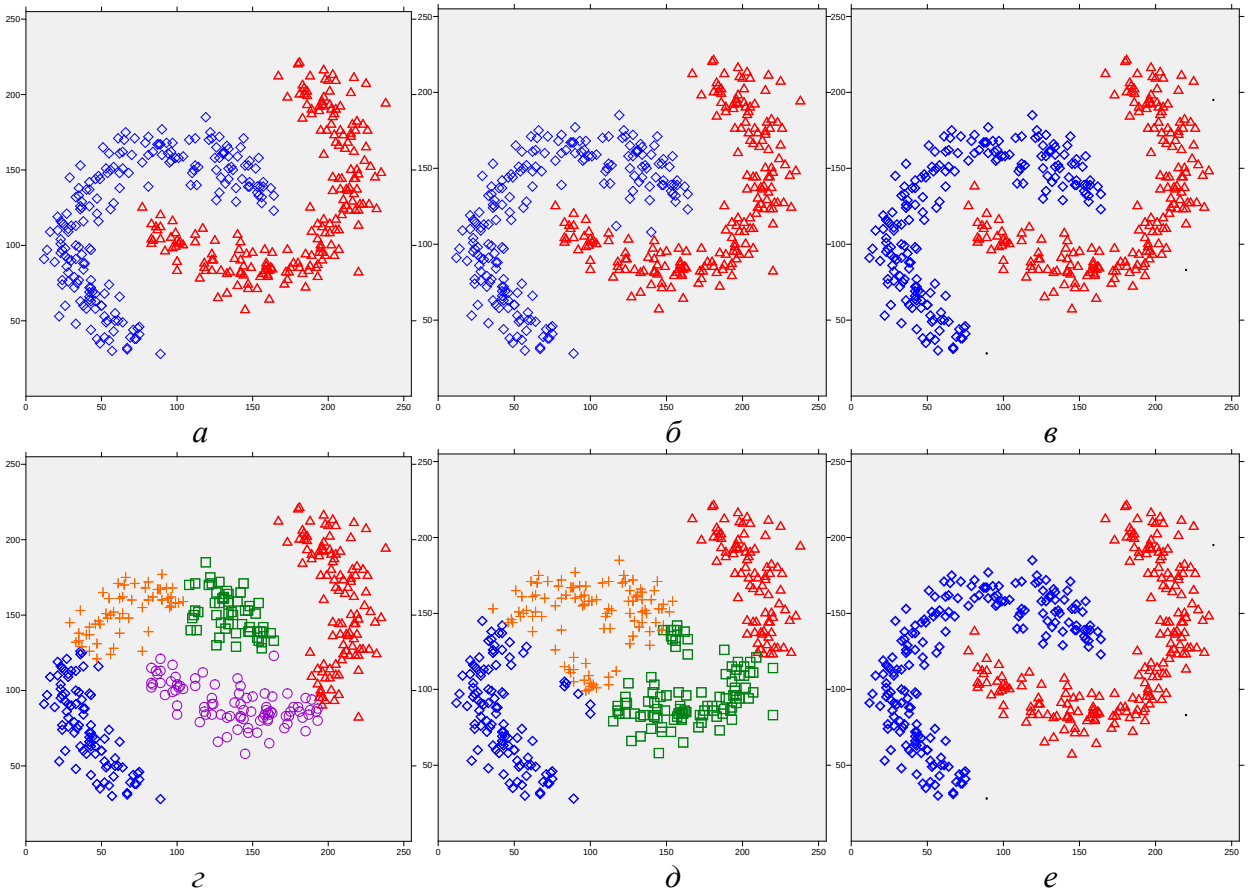
а *б* *в*

Рисунок 4.3 – Результаты обработки модельного набора данных 1: *а* – эталонное разбиение и результат выполнения алгоритмов MeanSC, EMeanSC, DBSCAN и SLINK;
б и *в* – результаты выполнения алгоритмов DENCLUE и k-средних соответственно



а *б* *в*

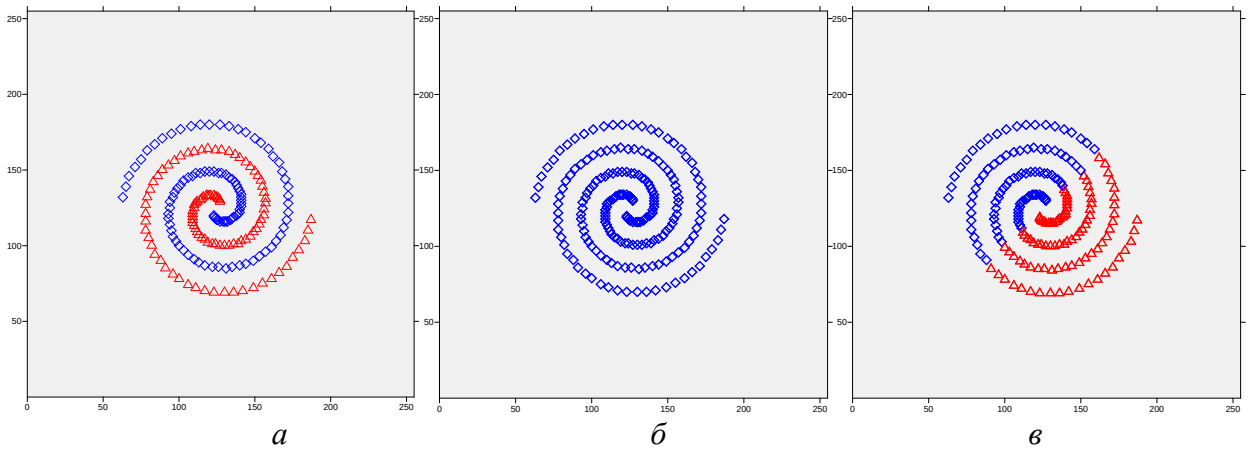
Рисунок 4.4 – Результаты обработки модельного набора данных 2: *а* – эталонное разбиение и результат выполнения алгоритмов MeanSC, EMeanSC, DBSCAN и SLINK; *б* и *в* – результаты выполнения алгоритмов DENCLUE и *k*-средних соответственно



а *б* *в*

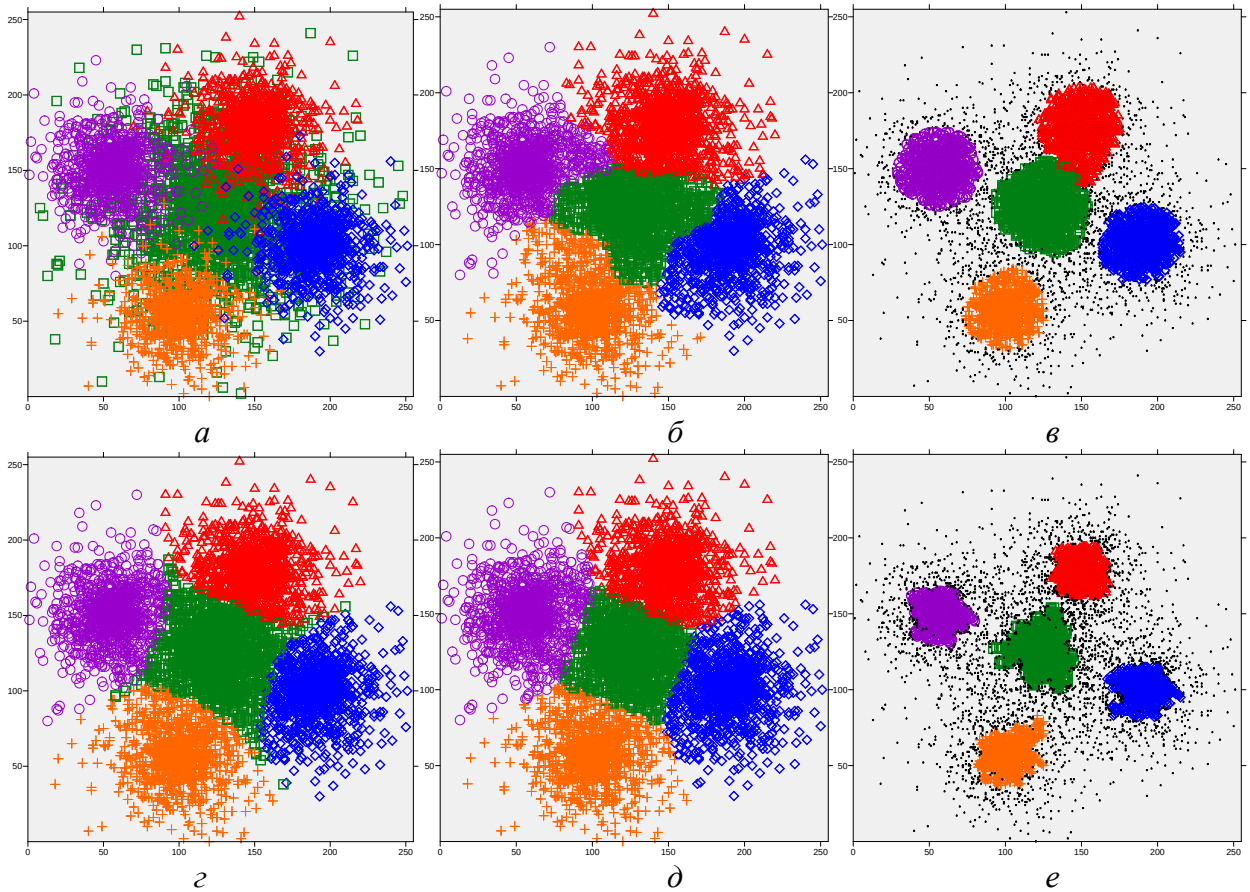
г *д* *е*

Рисунок 4.5 – Результаты обработки модельного набора данных 4: *а* – эталонное разбиение; *б* – результаты выполнения алгоритмов MeanSC и EMeanSC; *в*-*е* – результаты выполнения алгоритмов DBSCAN, DENCLUE, *k*-средних и SLINK соответственно. Точками на результате выполнения алгоритма SLINK отмечены кластеры размером менее 30 точек



а *б* *в*

Рисунок 4.6 – Результаты обработки модельного набора данных 3: *а* – эталонное разбиение и результат выполнения алгоритмов MeanSC, EMeanSC, DBSCAN и SLINK; *б* и *в* – результаты выполнения алгоритмов DENCLUE и *k*-средних соответственно



а *б* *в*

г *д* *е*

Рисунок 4.7 – Результаты обработки модельного набора данных 5: *а* – эталонное разбиение; *б* – результаты выполнения алгоритмов MeanSC и EMeanSC; *в*-*е* – результаты выполнения алгоритмов DBSCAN, DENCLUE, *k*-средних и SLINK соответственно. Точками на результате выполнения алгоритма SLINK отмечены кластеры размером менее 30 точек

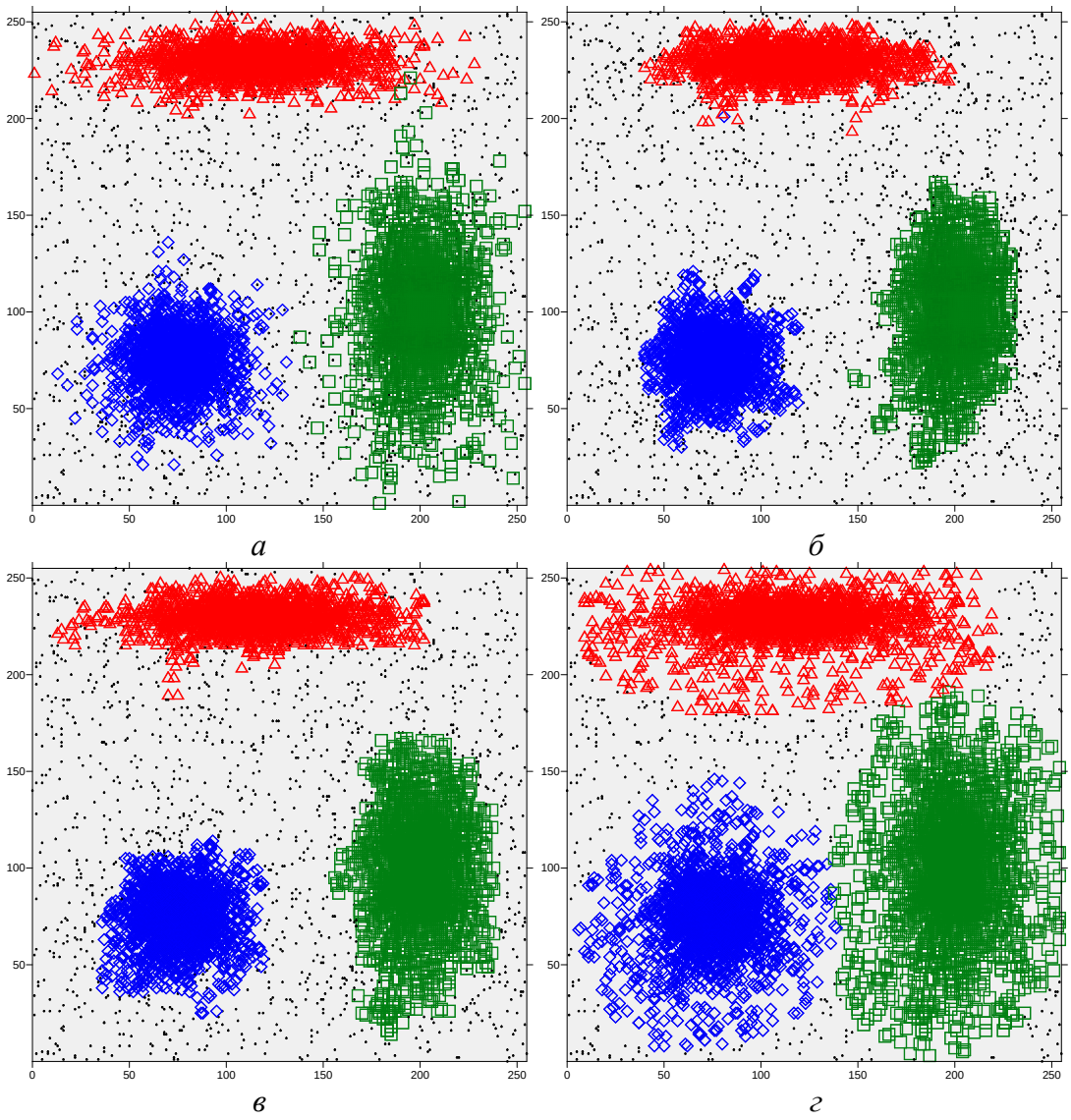


Рисунок 4.8 – Результаты обработки модельного набора данных 6: *а* – эталонное разбиение; *б-г* – результаты выполнения алгоритмов MeanSC, EMeanSC и DBSCAN

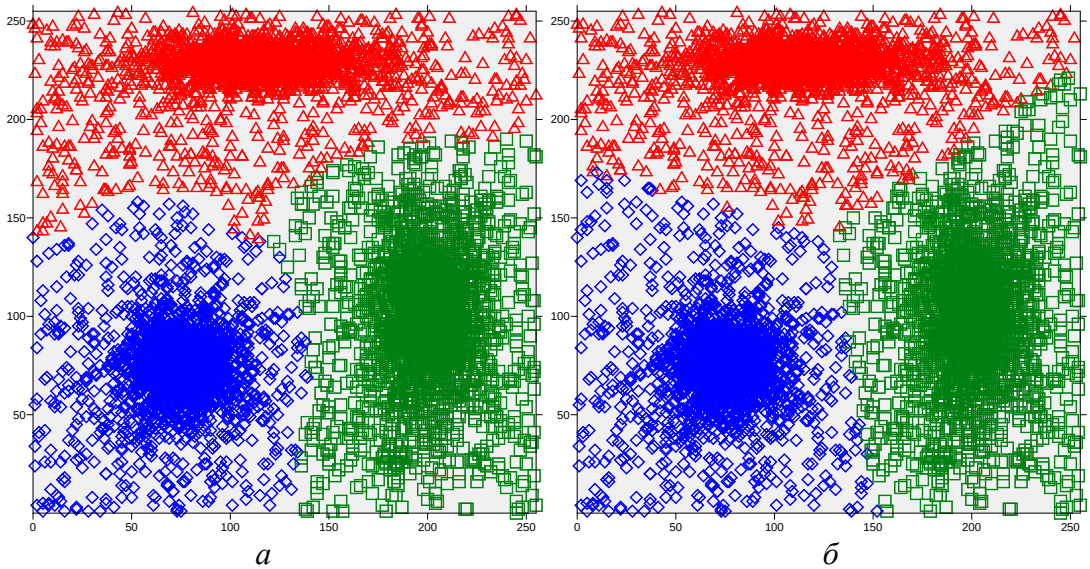
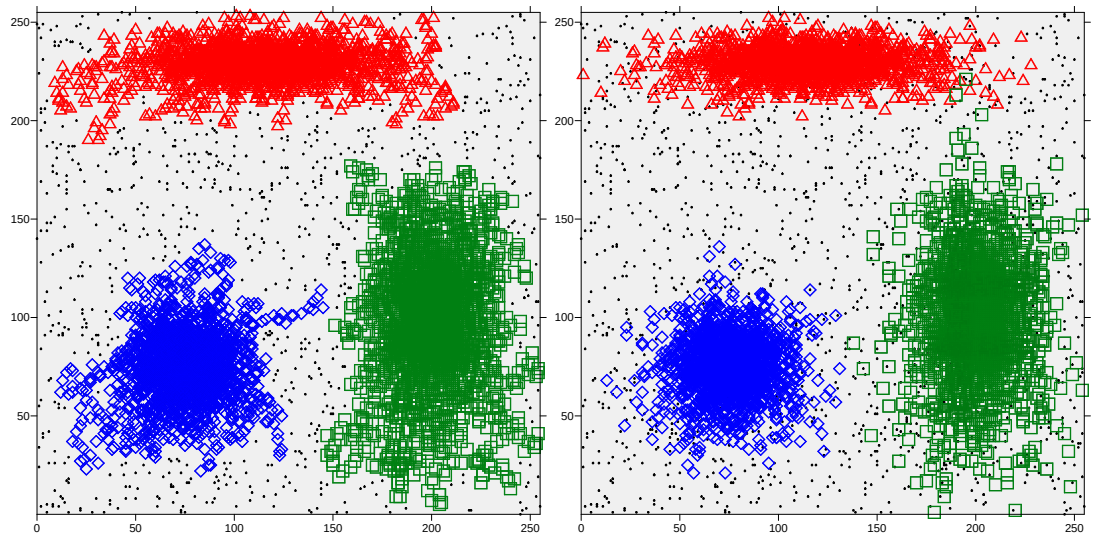
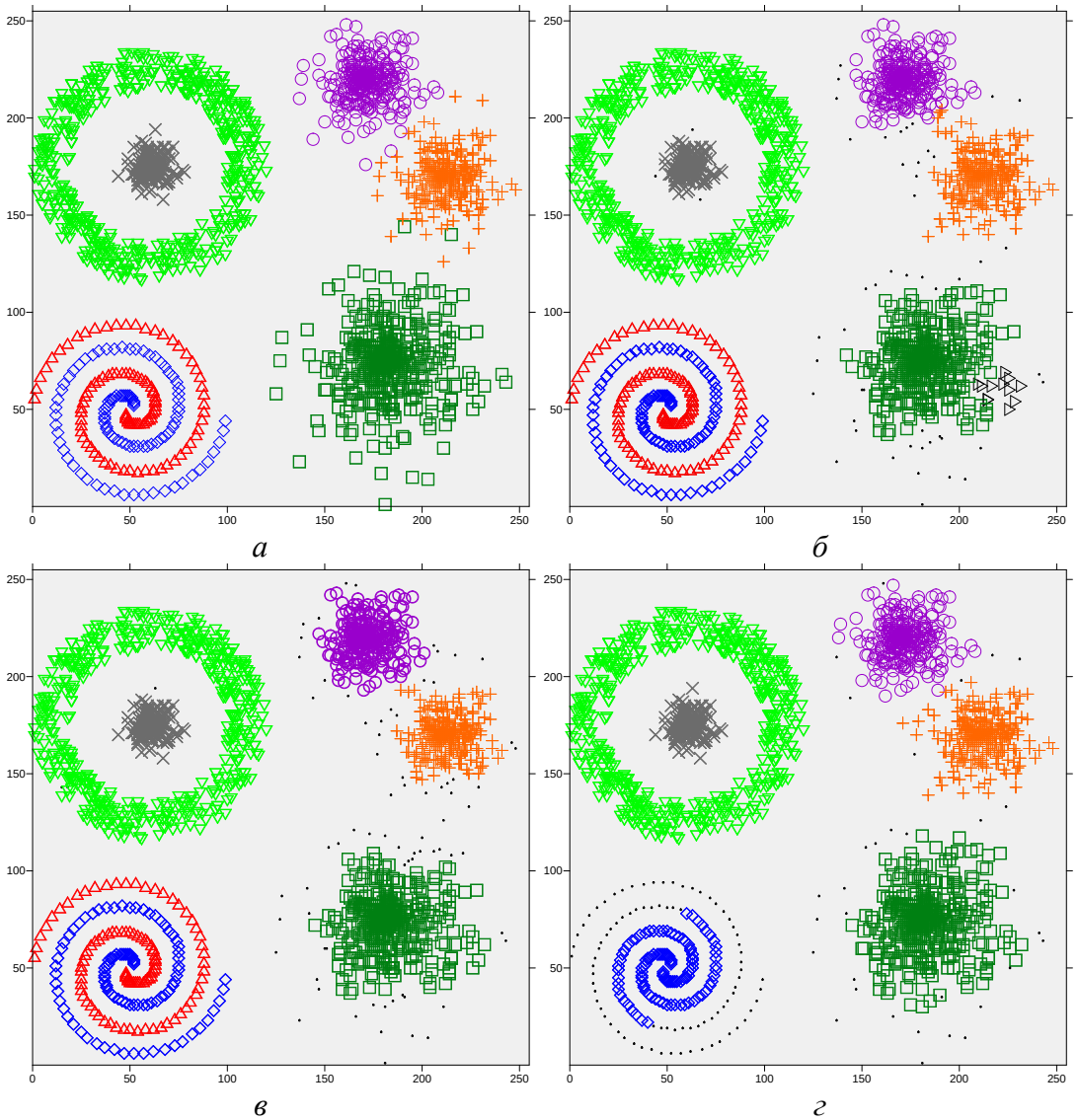


Рисунок 4.9 – Результаты обработки модельного набора данных 6: *а, б* –результаты выполнения алгоритмов *k*-средних и DENCLUE соответственно



а *б*

Рисунок 4.10 – Результаты обработки модельного набора данных 6: *а* – эталонное разбиение; *б* –результат выполнения алгоритма SLINK



а *б*

в *г*

Рисунок 4.11 – Результаты обработки модельного набора данных 7: *а* – эталонное разбиение; *б-г* – результаты выполнения алгоритмов MeanSC, EMeanSC и DBSCAN

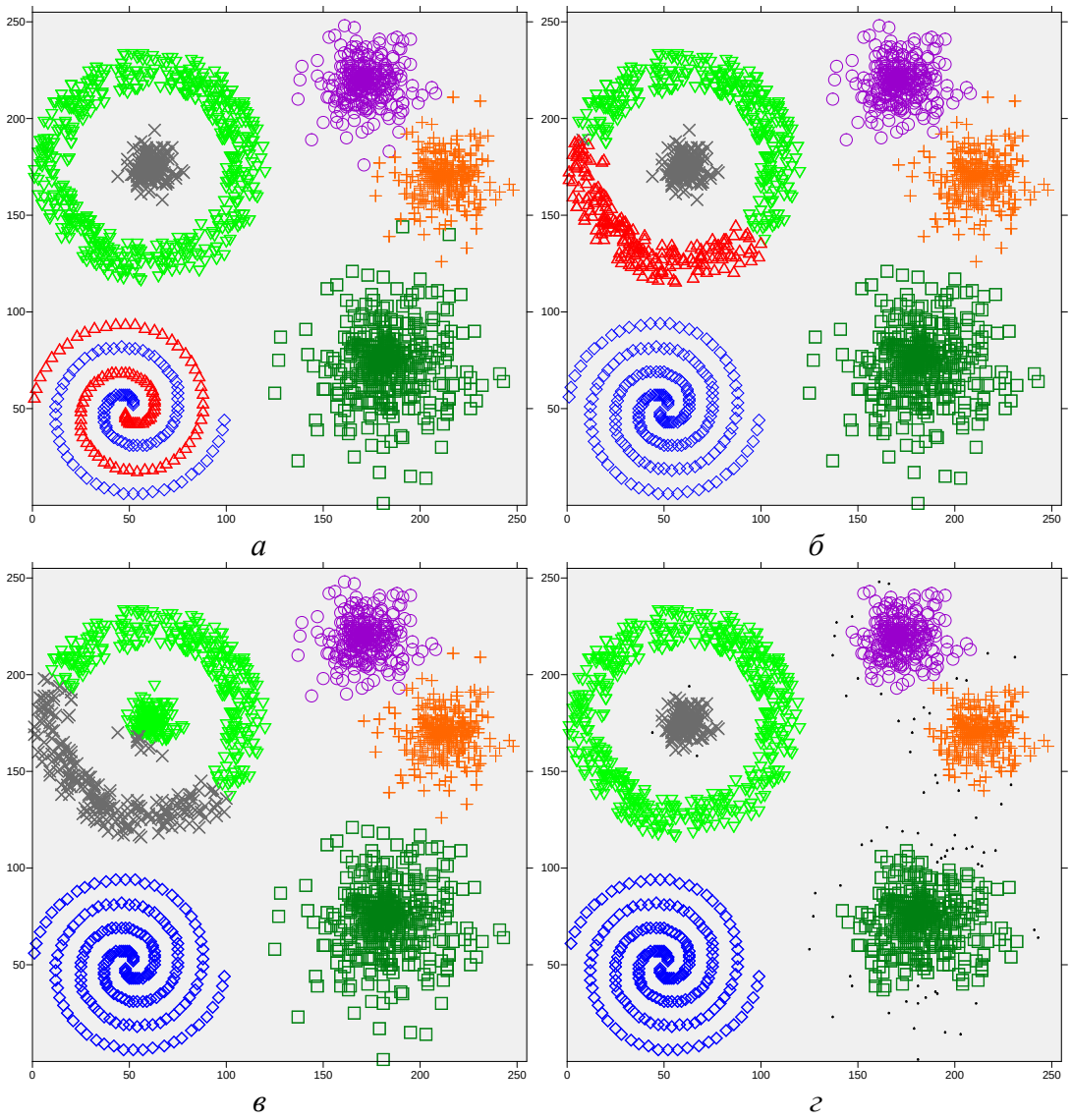


Рисунок 4.12 – Результаты обработки модельного набора данных 7: *а* – эталонное разбиение; *б-г* – результаты выполнения алгоритмов DENCLUE, *k*-средних и SLINK

Эксперименты на модельных наборах данных показали, что предложенные в рамках диссертационной работы алгоритмы позволяют получить результаты, не уступающие по качеству кластеризации лучшим непараметрическим алгоритмам, программные реализации которых можно найти в сети Интернет.

4.2. Экспериментальное исследование на реальных изображениях

Для оценки быстродействия алгоритмов использовалось восемь тестовых изображений – пять цифровых фотографий (рисунок 4.13) и три спутниковых изображения (рисунок 4.14). Обработка цифровых фотографий выполнялась по трём

каналам (в цветовом пространстве $R \times G \times B$), а спутниковых изображений – по четырём (красный, зелёный, синий и ближний инфракрасный).

Выполнялось сравнение предложенных алгоритмов с алгоритмами k-средних и ISODATA, которые включены во многие пакеты для обработки спутниковых изображений и поэтому часто используются при решении практических задач. В экспериментах использованы программные реализации алгоритмов из пакета ENVI, количество итераций было ограничено 10. Центры кластеров, полученные на предыдущей итерации, использовались для инициализации следующей, что позволило получить более качественные результаты. Кроме того, в экспериментах использованы программные реализации наиболее эффективных плотностных алгоритмов DBSCAN, OPTICS, DENCLUE и MeanShift из пакетов ELKI и Smile. В экспериментах количество итераций «среднего сдвига» для алгоритма MeanShift было ограничено 10.

Результаты эксперимента представлены в таблице 4.2. Прочерки в таблице соответствуют неприемлемо высокому времени обработки (более 18 часов).

Рисунок 4.13 – Тестовые изображения 1–5 (цифровые фотографии). Размер изображений составляет 0.3, 1, 2.2, 5 и 13.8 млн пикселей соответственно

Рисунок 4.14 – Тестовые изображения 6–8 (фрагменты снимков, полученных со спутника WorldView-2). Размер изображений составляет 4.2, 9 и 12 млн пикселей соответственно

Для оценки эффективности распараллеливания с использованием стандарта OpenMP выполнено сравнение времени обработки тестовых изображений с использованием последовательной (однопоточной) и параллельной (многопоточной) версий предложенных алгоритмов. Результаты приведены в таблице 4.3. Видно, что применение параллельных вычислений позволило существенно уменьшить время работы, особенно при обработке изображений большого размера.

Таблица 4.2 – Результаты обработки тестовых изображений (время в секундах)

Изображение	1	2	3	4	5	6	7	8
Размер (млн пикселей)	0.3	1	2.2	5	13.8	4.2	9	12
Число каналов	3	3	3	3	3	4	4	4
MeanSC	0.09	0.51	0.86	1.44	8.99	1.44	8.16	4.2
EMeanSC	0.39	2.25	3.16	5.21	31.31	4.97	28.74	10.47
MeanShift	2.91	52	102	67	388	4138	217	62388
k-средних	0.5	36	5	17	1196	75	302	588
ISODATA	1	15	5	9	1178	68	332	337
DBSCAN	194	2731	13098	–	–	39965	–	–
OPTICS	638	5244	40013	–	–	–	–	–
DENCLUE	6934	39849	–	–	–	–	–	–

Таблица 4.3 – Сравнение последовательной и параллельной версий предложенных алгоритмов (время в секундах)

Изображение	Размер (млн пикселей)	Число каналов	Время обработки (1 поток), с				Время обработки (8 потоков), с			
			MeanSC			EMeanSC	MeanSC			EMeanSC
			$m = 11$	$m = 13$	$m = 16$		$m = 11$	$m = 13$	$m = 16$	
1	0.3	3	0.16	0.11	0.1	0.4	0.11	0.1	0.09	0.39
2	1.0	3	1.14	1	0.64	3.28	0.61	0.56	0.51	2.25
3	2.2	3	2.69	2.15	2.18	8.08	0.98	0.86	0.86	3.16
4	5.0	3	5.48	4.09	3.06	13.63	1.93	1.71	1.44	5.21
5	13.8	3	38.03	28.31	20.53	87.82	12	9.93	8.99	31.31
6	4.2	4	2.91	2.52	2.21	7.89	1.72	1.64	1.44	4.97
7	9.0	4	48.21	31.31	18.75	99.16	11.65	8.4	8.16	28.74
8	12.0	4	4.46	4.53	4.42	14.75	3.11	2.73	4.2	10.47

Анализ результатов показывает, что алгоритмы, включенные в пакеты ELKI и Smile, не позволяют оперативно обрабатывать изображения высокого разрешения. Кроме того, время их работы значительно увеличивается с ростом числа каналов. Алгоритмы из пакета ENVI лучше адаптированы к анализу изображений, однако время обработки изображений размером более 9 млн пикселей превышает 5 минут. Алгоритмы, предложенные в рамках диссертационной работы, позволяют выполнять сегментацию мультиспектральных изображений большого размера в диалоговом режиме.

Выводы по главе

1. Выполнен сравнительный анализ алгоритмов MeanSC и EMeanSC с алгоритмами, включёнными в распространённый пакет для обработки спутниковых данных ENVI и в пакеты для анализа данных ELKI и Smile.
2. На модельных данных показано, что алгоритмы MeanSC и EMeanSC превосходят известные непараметрические алгоритмы по качеству кластеризации и/или времени обработки.

3. На реальных изображениях продемонстрированно, что разработанные алгоритмы позволяют обрабатывать мультиспектральные изображения размером до 14 млн пикселей в диалоговом режиме.
4. Продемонстрирована эффективность параллельных реализаций предложенных алгоритмов.

ГЛАВА 5. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ НА ОСНОВЕ РАЗРАБОТАННЫХ АЛГОРИТМОВ И РЕШЕНИЕ ПРАКТИЧЕСКИХ ЗАДАЧ

В настоящее время решение практических задач, связанных с обработкой пространственных данных, производится с помощью традиционных, но, зачастую, устаревших, методов анализа, включённых в состав автономных программных пакетов. Поэтому одной из целей диссертационной работы являлось создание схемы интеграции, обеспечивающей разработчикам простой и быстрый механизм внедрения новых и обновления существующих алгоритмов обработки пространственных данных, а потенциальным пользователям – прозрачный доступ к ним.

В ходе выполнения диссертационной работы разработаны два эффективных механизма внедрения алгоритмов обработки пространственных данных: в виде стандартизованных веб-сервисов и в качестве модулей для открытой геоинформационной системы GRASS GIS. Кроме того, создан пакет программ «Image Processing Toolkit», предназначенный для сегментации мультиспектральных изображений и включающий набор эффективных алгоритмов кластеризации, разработанных сотрудниками Института вычислительных технологий СО РАН в рамках различных проектов и грантов.

5.1. Платформа для предоставления алгоритмов обработки пространственных данных в виде веб-сервисов

Сложность современных пакетов обработки спутниковых снимков, их существенная стоимость и необходимость постоянного обновления значительно затрудняют их широкое использование рядовым потребителем.

Начиная с 2007 года, консорциум OGC разрабатывает протокол предоставления сервисов обработки пространственных данных WPS (Web Processing Service) [11]. Предоставление алгоритмов обработки в виде веб-сервисов (WPS-процессов) позволит значительно упростить внедрение передовых технологий и предоставить пользователям доступ к распределённому хранилищу современных алгоритмов.

В настоящее время активно развиваются программные системы с открытым исходным кодом, предоставляющие инструментарий для реализации WPS-процессов. Использование WPS-процессов конечным пользователем возможно при помощи ГИС-пакетов, в которых реализована поддержка этого протокола. На данный момент к ним относятся не только открытые ГИС (такие, как uDig, QGIS, ILWIS и др.), но и коммерческий пакет ArcGIS. Кроме того, существуют библиотеки классов с открытым исходным кодом для реализации WPS-клиента на языках программирования Java и JavaScript.

5.1.1. Технология внедрения алгоритмов

Спецификация WPS описывает стандартный интерфейс для публикации процессов обработки пространственных данных, а также правила поиска и доступа к ним со стороны клиента. Процессом может являться любой алгоритм (или численная модель), использующий пространственно скоординированные данные. Под публикацией понимается предоставление стандартизованной информации, необходимой для доступа к процессу, и метаданных на естественном языке, позволяющих осуществлять поиск и использование процесса.

В ходе выполнения диссертационной работы создана система веб-сервисов (рисунок 5.1), доступная по адресу http://wps.ict.nsc.ru:8080/wps/WebProcessingService (протокол доступа – WPS). Реализация выполнена на платформе Java 1.6, что позволило обеспечить платформенную независимость. Ядром системы является WPS-сервер, который создан в рамках проекта 52°North (https://52north.org/software/software-projects/wps) и представляет собой веб-приложение, работающее под управлением контейнера сервлетов Apache Tomcat. Он осуществляет интерпретацию входных и выходных данных согласно спецификации протокола WPS и выполняет функции контейнера для неограниченного числа WPS-процессов. Кроме того, он обеспечивает доступ к части функционала GRASS GIS.

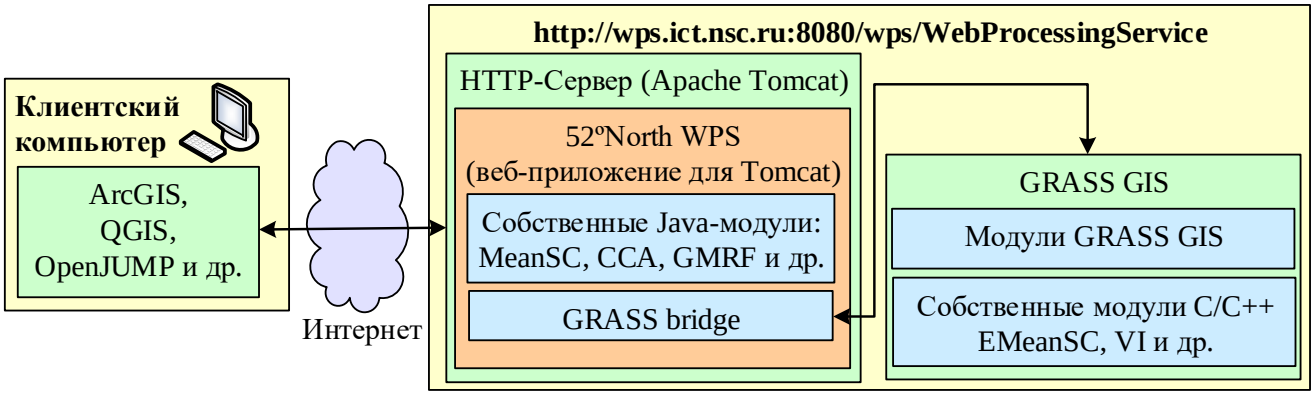


Рисунок 5.1 – Структурная схема разработанной системы сервисов

Для обработки данных с помощью WPS-процесса пользователь вводит в клиентском приложении адрес WPS-сервера, после чего ему предоставляется список доступных процессов и их описания (метаданные на естественном языке). Выбрав необходимый алгоритм, пользователь указывает значения входных и выходных параметров в соответствии со спецификацией протокола WPS. Например, для алгоритмов классификации входными параметрами являются классифицируемое растровое изображение, обучающая выборка (для обучаемых и полуобучаемых алгоритмов), а также набор параметров, специфичных для конкретного алгоритма. Значениями параметров могут быть как данные, находящиеся на компьютере пользователя, так и результаты выполнения запросов к удалённым WPS/WMS/WFS-серверам [9-11]. В этом случае запрос обрабатывается распределённо, без необходимости сохранения промежуточных результатов.

В настоящее время в виде WPS-процессов опубликовано пять эффективных непараметрических алгоритмов, созданных в ИВТ СО РАН в рамках различных проектов и грантов. Они позволяют решать широкий круг задач, связанных с распознаванием образов и анализом мультиспектральных спутниковых данных. Помимо алгоритма кластеризации MeanSC, опубликованного в ходе выполнения диссертационной работы, по протоколу WPS доступны алгоритмы автоматической классификации CCA [146] и ECCA [116, 117], а также непараметрический иерархический классификатор для обработки данных дистанционного зондирования [147] и алгоритм классификации с полуобучением [148].

Алгоритм кластеризации CCA разработан в рамках комбинации плотностного и сеточного подходов и позволяет выделять многомодовые кластеры сложной формы. Варьирование значения специального параметра приводит к получению результатов различной степени подробности. Быстродействие алгоритма позволяет проводить обработку данных в диалоговом режиме.

Ансамблевый алгоритм ECCA разработан на основе алгоритма CCA. Использование ансамблевого подхода позволяет значительно снизить влияние размера сетки на качество кластеризации.

Непараметрический иерархический классификатор. Традиционный подход к построению непараметрических алгоритмов, основанных на оценках Розенблатта – Парзена, заключается в подстановке в байесовское решающее правило вместо неизвестных вероятностных характеристик классов соответствующих им оценок, полученных по обучающим выборкам. Вычисление таких оценок является трудоёмкой операцией. Для повышения её производительности применяется переход к пространству признаков меньшей размерности (извлечение признаков). Существуют эффективные (в смысле вероятности ошибки классификации) методы извлечения информативных признаков, обеспечивающие хорошие результаты в двухклассовом случае, но с ростом числа классов информативность выделяемых признаков существенно снижается.

Опубликованный алгоритм обеспечивает хорошие результаты как в двухклассовом, так и в многоклассовом случае за счет того, что решение общей задачи сводится к решению нескольких задач с меньшим числом классов благодаря введению иерархии классов. Для каждой подзадачи определяется соответствующий набор информативных признаков.

Алгоритм классификации с полуобучением. В задачах обучаемой классификации спутниковых изображений процесс получения обучающей выборки, необходимой для построения решающего правила, зачастую связан со значительными материальными и временными затратами. Поэтому на практике обучающая выборка, как правило, имеется лишь для интересующих пользователя классов и при этом является непредставительной.

В то же время при классификации изображений всегда доступен большой объём непомеченных данных. В этих условиях для расширения обучающей выборки можно использовать методы классификации с полуобучением. Они позволяют использовать информацию о плотности распределения, содержащуюся в непомеченных данных.

Опубликованный алгоритм позволяет получить представительную обучающую выборку из исходной на основе анализа непомеченных данных. В результате получается выборка, достаточная для применения алгоритма классификации с обучением (в данном случае использован классификатор Розенблатта – Парзена с нормальным ядром).

Разработанные в Институте вычислительных технологий СО РАН и предоставляемые в виде веб-сервисов непараметрические алгоритмы классификации позволяют решать широкий круг задач, связанных с обработкой и анализом мультиспектральных спутниковых изображений.

Публикация алгоритмов обработки пространственных данных в виде веб-сервисов позволяет любому потенциальному пользователю на основе выбранного клиентского приложения (с удобным и привычным интерфейсом) и программно-алгоритмических ресурсов системы создать ГИС, идеально подходящую для решения конкретной задачи и имеющую доступ к необходимым вычислительных (кластерные системы ИВТ СО РАН, Новосибирского государственного университета и др.) и информационным (пространственные данные, предоставляемые в виде веб-сервисов) ресурсам.

5.1.2. Схема интеграции WPS-процессов в распределённую сервис-ориентированную геоинформационную систему ИВТ СО РАН

В Институте вычислительных технологий СО РАН, начиная с 2006 г., активно развивается сервис-ориентированная геоинформационная система [149], которая создана на основе каталога спутниковых данных Новосибирского научного центра СО РАН (http://gis-app.ict.nsc.ru/catalogue) [150].

Основная цель разработки системы – организация единой точки доступа к разнородным, географически распределённым хранилищам пространственных данных и средствам для их визуализации, анализа и обработки на современных вычислительных комплексах. Кроме того, система позволит обеспечить возможность оперативного взаимодействия разрозненных групп исследователей (зачастую, разделённых географически) при выполнении различных проектов и грантов. Особо остро потребность в системах такого рода ощущается при проведении интеграционных исследований, когда решаемые задачи требуют привлечения учёных из различных областей науки.

Разрабатываемая система работает под управлением операционной системы Linux (дистрибутив PUIAS Linux) и базируется на наборе стандартных и специализированных программных продуктов с открытым исходным кодом, распространяемых по лицензии GPL (GNU General Public License). Она полностью удовлетворяет требованиям консорциума по стандартизации OGC (Open Geospatial Consortium), предъявляемым к геоинформационным системам.

Доступ к системе осуществляется двумя способами: через веб-браузер (что обеспечивает платформенную независимость) и по стандартизованным протоколам предоставления пространственных данных и алгоритмов WMS/WFS/WPS (что позволяет работать с ней, используя любые удобные пользователю коммерческие или открытые настольные ГИС).

Общая структурная схема системы представлена на рисунке 5.2. Центральным её блоком является подсистема картографических сервисов [14], реализованная на основе пакета GeoServer (http://geoserver.org). Подсистема обеспечивает доступ к картографической информации, хранимой в системе (базовые подложки, векторные слои, построенные по базам данных, и др.), в виде веб-сервисов.
Разработанная в рамках диссертационной работы платформа является составной частью подсистемы сервисов (рисунок 5.3) и предназначена для расширения функциональности системы. Она позволяет интегрировать алгоритмы обработки спутниковых данных, созданные в ИВТ и других институтах СО РАН и предоставлять их широкому кругу потенциальных пользователей в виде веб-сервисов.

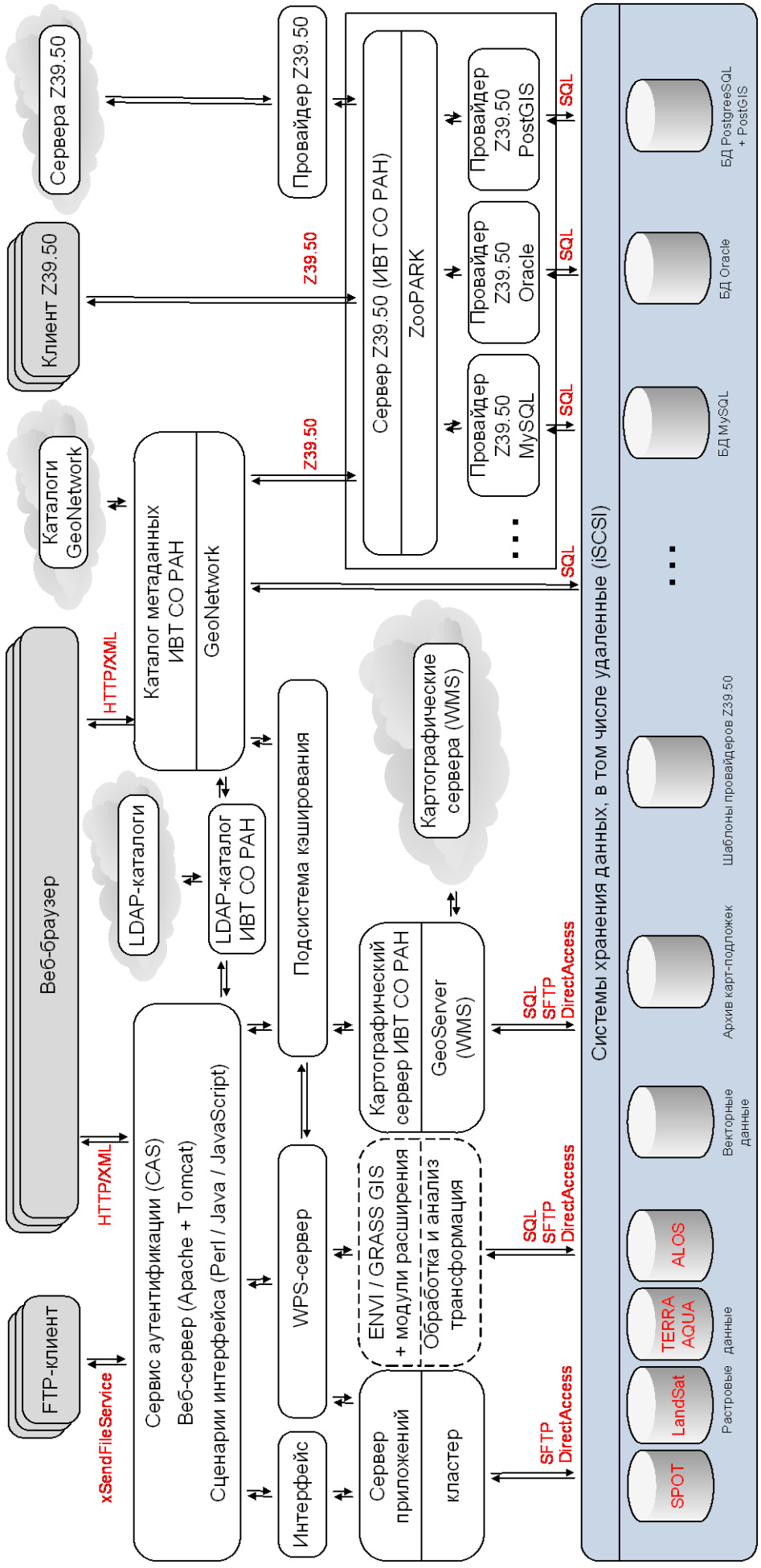


Рисунок 5.2 – Структура сервис-ориентированной геоинформационной системы ИВТ СО РАН

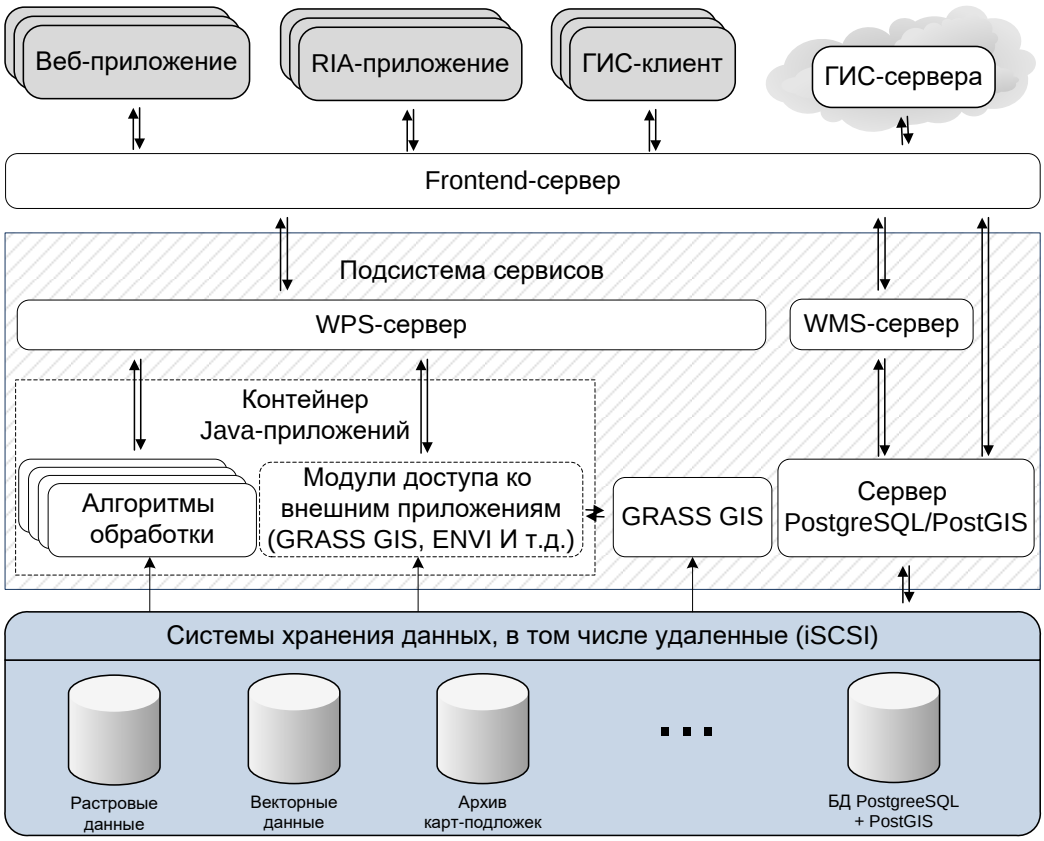


Рисунок 5.3 – Структурная схема подсистемы сервисов

Для обеспечения эффективного поиска по базам метаданных в систему включены модули поддержки протокола Z39.50 [13]. Поисковая подсистема позволяет не только находить данные по метаданным, но и выполнять комплексные запросы (содержащие географические координаты, дату и время съёмки, параметры облачности и др.).

5.2. Внедрение в виде модулей в открытую геоинформационную систему GRASS GIS

При проведении исследований в полевых условиях часто возникает необходимость обработки пространственных данных при низкой скорости или отсутствии подключения к сети Интернет. В таких условиях разработанная система сервисов является бесполезной, поэтому в рамках диссертационной работы также исследовалась возможность интеграции новых алгоритмов в виде модулей в одну из существующих геоинформационных систем (ГИС).

К настоящему времени разработано достаточно много ГИС общего назначения, как коммерческих (ITTVIS ENVI, ESRI ArcGIS, ERDAS Imagine, MapInfo и

др.), так и свободно распространяемых (GRASS GIS, SAGA GIS, uDig, QGIS, ILWIS и др.).

Коммерческие пакеты ориентированы на встроенный функционал, расширяемый за счёт отдельно приобретаемых модулей (стоимость которых иногда сравнима со стоимостью самой ГИС). В коммерческие ГИС достаточно сложно встроить свои алгоритмы, зачастую для этого необходимо освоить встроенный в пакет язык программирования, документация к которому также не является бесплатной. Функционал свободно распространяемых пакетов, наоборот, практически полностью (за исключением модулей ввода/вывода и визуализации) опирается на вебсервисы и модули расширения, процесс написания которых предельно упрощён, детально описан в свободно распространяемой документации и не требует изучения узкоспециализированных языков программирования.

Исключением является пакет GRASS GIS (http://grass.osgeo.org), сочетающий достоинства коммерческих и свободно распространяемых геоинформационных систем. Его отличительные особенности – полная интеграция в среду UNIX, поддержка основных типов пространственных данных, модульность, мощный процессор обработки растровых данных и наличие открытого инструментария для быстрой и эффективной разработки модулей расширения. Использование библиотек GDAL (http://www.gdal.org) и PROJ (https://trac.osgeo.org/proj) обеспечивает поддержку всех современных стандартов геоданных и большой набор функций для трансформации и перепроецирования изображений. По функциональности GRASS GIS не уступает коммерческим аналогам. Он позволяет разрабатывать модули расширения практически на всех языках программирования, для которых есть компилятор под UNIX (C/C++, Java, Fortran, Perl, sh и др.). Пакет допускает выполнение ресурсоемких алгоритмов на высокопроизводительных вычислительных системах. К нему подключены библиотеки для работы практически со всеми современными системами управления базами данных. Ядро GRASS GIS является функциональным расширением командной оболочки UNIX, что позволяет автоматизировать любой процесс обработки, не требующий активного участия пользователя. Кроме того, GRASS GIS имеет возможность работы в многопользовательском режиме.

Помимо 52°North, в настоящее время активно развиваются два пакета программ с открытым исходным кодом, обеспечивающие доступ к большей части функционала GRASS GIS по протоколу WPS: PyWPS (http://pywps.wald.intevation.org) и ZOO Project (http://zoo-project.org). Это даёт возможность создать на основе GRASS GIS набор веб-сервисов, позволяющий легко дополнить произвольную коммерческую или свободно распространяемую ГИС, поддерживающую протокол WPS, функционалом, необходимым для решения любой отдельно взятой задачи.

Учитывая перечисленные особенности, на данный момент GRASS GIS является наиболее перспективным пакетом для внедрения современных алгоритмов обработки и анализа пространственных данных.

Спутниковые изображения и другие пространственные данные представляются в GRASS GIS в виде информационных слоёв (мультиспектральные изображения – в виде набора слоёв, по слою на каждый спектральных канал). Для работы необходимо создать набор, объединяющий информационные слои, которые будут использованы при решении задачи. При добавлении в набор, все информационные слои автоматически приводятся к одной картографической проекции и одному пространственному разрешению, из них вырезается только интересующий регион. Это значительно упрощает дальнейшую обработку, позволяя единообразно работать со всеми имеющимися пространственными данными.

В рамках диссертационной работы на базе разработанного ансамблевого алгоритма EMeanSC создан модуль i.ict.emeansc для GRASS GIS. Реализация модуля выполнена на языке программирования C++ под UNIX. Для выполнения обработки необходимо задать следующие параметры (на примере модуля r.ict.emeansc):

- параметры модуля GRASS GIS:
  - набор входных слоёв, разделённых запятыми (параметр *input*);
  - выходной слой (параметр *output*);
  - разрешение перезаписывать существующие файлы (флаг *overwrite*);
  - необходимость выведения детального отчёта (флаг *verbose*);
- параметры алгоритма EMeanSC:

- набор значений параметра сглаживания $\bar{m}$ для формирования ансамбля, разделённых запятыми (параметр m);
- минимальная плотность ε (параметр eps);
- порог объединения для компонент связности T (параметр T);
- порог объединения для построения дендрограммы T_d (параметр Td).

Задание параметров осуществляется либо посредством динамически генерируемого средствами GRASS GIS графического интерфейса (рисунок 5.4), либо из командной строки Unix. Например, команда

r.ict.emeansc $--$**overwrite** $--$**verbose** \

input = **L.3@USER**,**L.4@USER**,**L.5@USER** \

output = **NEWMAP@USER m** = **10**,**15**,**20 eps** = **100 T** = **0.3 Td** = **0.6**

приведёт к сегментации мультиспектрального изображения, сформированного из спектральных слоёв L.3, L.4 и L.5 пользовательского набора USER, алгоритмом EMeanSC с параметрами $\bar{m} = \{10, 15, 20\}$, $\varepsilon = 100$, $T = 0.3$, $T_d = 0.6$. Результат будет записан в слой NEWMAP набора USER (если слой существует, он будет перезаписан, если не существует – создан). После обработки пользователю будет предоставлен детальный отчёт.

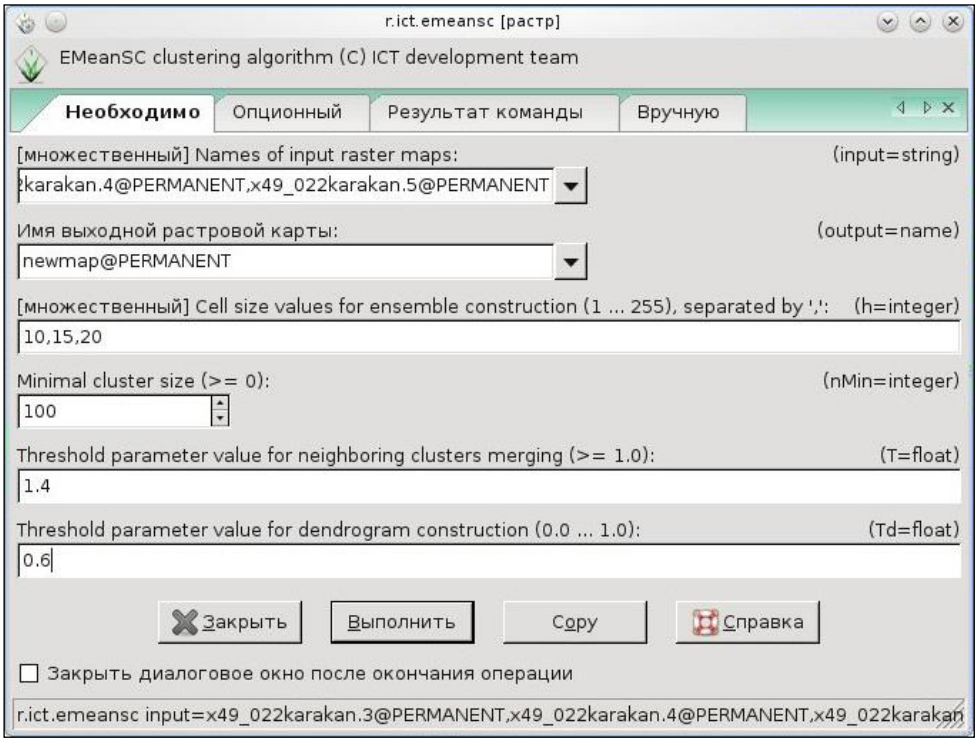


Рисунок 5.4 – Графический интерфейс ввода параметров модуля i.ict.emeansc

Стоит заметить, что возможность запуска модуля из командной строки UNIX позволяет формировать цепочки алгоритмов для частичной автоматизации процессов обработки и анализа данных.

5.3. Пакет прикладных программ для обработки мультиспектральных изображений «Image Processing Toolkit»

В рамках диссертационной работы был создан пакет программ «Image Processing Toolkit», предназначенный для обработки мультиспектральных спутниковых изображений с целью изучения и оценки состояния природных объектов. Пакет реализован на языке программирования C++ в среде Microsoft Visual Studio 2019 и работает под управлением операционной системы Windows. Он базируется на библиотеке классов MFC и включает набор эффективных алгоритмов классификации. Пакет допускает расширение за счёт дополнительных функциональных модулей. Поиск и загрузка модулей расширения производится автоматически, в момент запуска приложения.

5.3.1. Структура и основные функции пакета

Пакет «Image Processing Toolkit» включает четыре основных модуля. Схема взаимодействия и функции модулей представлены на рисунке 5.5.

Модуль Algolib является ядром пакета. Он обеспечивает следующую функциональность:

- загрузка и использование дополнительных модулей;
- загрузка, преобразование во внутренний формат и сохранение модельных данных в специальном формате;
- загрузка, преобразование во внутренний формат и сохранение растровых данных в общепринятых форматах;
- запуск имеющихся алгоритмов (в т.ч. из модулей расширения);
- изменение цветовой палитры картосхемы;
- сохранение результатов классификации в общепринятых форматах.

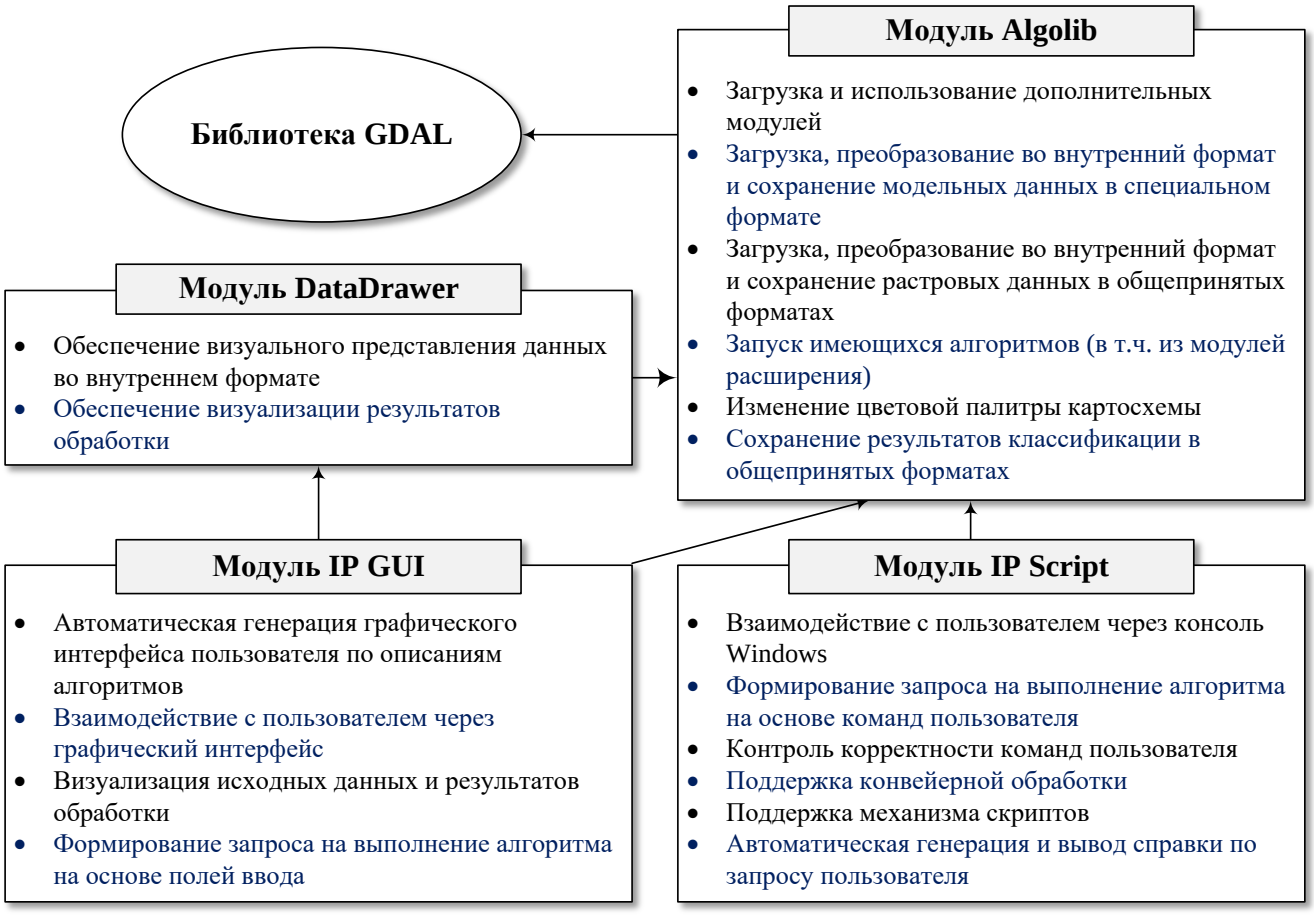


Рисунок 5.5 – Схема взаимодействия основных модулей пакета «Image Processing Toolkit»

Диаграмма классов модуля Algolib представлена на рисунке 5.6. Основные классы:

- AData – внутреннее представление данных (модуль расширения, обеспечивающий поддержку нового типа данных, должен включать класс, наследованный от AData);
- Algorithm – произвольный алгоритм обработки данных (модуль расширения, обеспечивающий поддержку нового алгоритма обработки, должен включать класс, наследованный от Algorithm);
- ROI – внутреннее представление области интереса, необходимой для поддержки алгоритмов классификации с обучением.

Модуль DataDrawer предназначен для визуализации данных, представленных во внутреннем формате, а также результатов обработки.

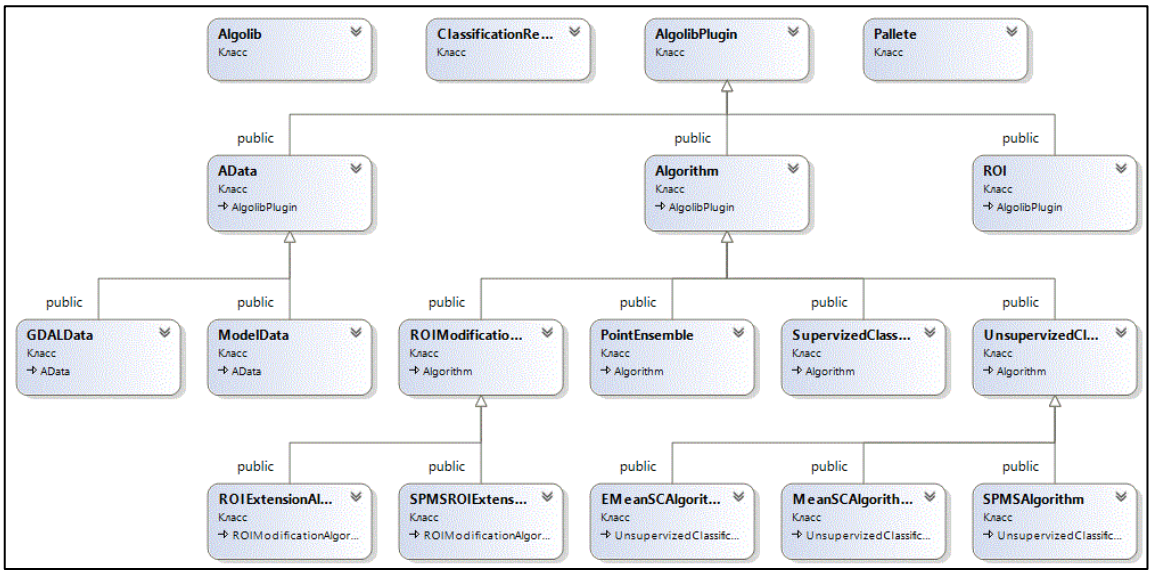


Рисунок 5.6 – Диаграмма классов модуля Algolib

Модуль IP GUI предназначен для взаимодействия с пользователем через графический интерфейс. Основные функции модуля:

- автоматическая генерация графического интерфейса пользователя в момент запуска приложения по описаниям алгоритмов, получаемым от модуля Algolib;
- взаимодействие с пользователем через графический интерфейс;
- загрузка данных и сохранение результатов обработки с помощью модуля Algolib;
- визуализация исходных данных и результатов обработки с помощью модуля DataDrawer;
- формирование запросов на обработку данных на основе полей ввода;
- запуск алгоритмов с помощью модуля Algolib.

Модуль IP Script предназначен для взаимодействия с пользователем посредством командной строки Windows. Основные функции модуля:

- взаимодействие с пользователем через консоль Windows;
- контроль корректности команд пользователя;
- формирование запроса на выполнение алгоритма на основе команд пользователя;
- поддержка конвейерной обработки (передача результатов выполнения алгоритма на вход следующему);

- поддержка механизма скриптов (возможность выполнения заранее сформированных последовательностей команд);
- автоматическая генерация и вывод на экран справки на основе описаний алгоритмов, получаемых от модуля Algolib.

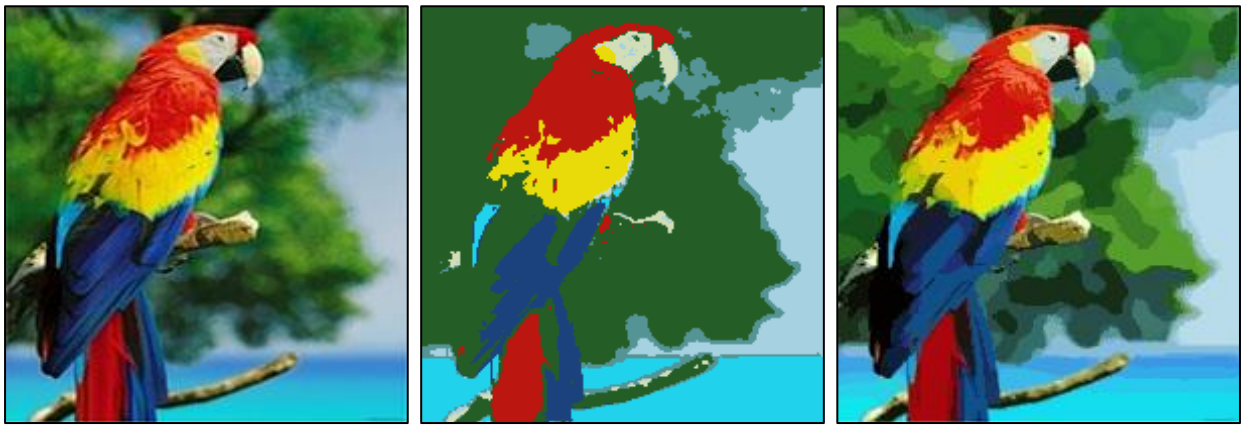
5.3.2. Алгоритмы, включённые в пакет «Image Processing Toolkit»

Помимо алгоритмов кластеризации, разработанных в рамках диссертационной работы, в пакет «Image Processing Toolkit» включены следующие алгоритмы.

Алгоритм сегментации изображений SPMeanSC. Алгоритм является модификацией алгоритма MeanSC, позволяющей учитывать пространственную информацию. Для этого в процедуре «среднего сдвига» при вычислении координат нового центра учитываются только пиксели, удалённые от текущего центра не более, чем на h_r в плоскости изображения и не более, чем на h_s в пространстве признаков. В качестве расстояния между пикселями используется евклидово расстояние между векторами (в пространстве признаков расстояние определяется по векторам спектральных яркостей, а в плоскости изображения – по векторам координат пикселей). После этого сегменты с близкими центрами (на расстоянии не более h_r в плоскости изображения и не более $h_s/2$ в пространстве признаков) объединяются и для каждого сегмента вычисляется представитель (среднее значение спектральных яркостей всех пикселей, отнесённых к сегменту). Для запуска алгоритма требуется два параметра – h_r и h_s.

Пример сегментации цветного изображения алгоритмами MeanSC (8 кластеров) и SPMeanSC (871 кластер) представлен на рисунке 5.7.

Алгоритм наращивания обучающей выборки SPMSROIExtension. В задачах классификации спутниковых изображений процесс получения обучающей выборки (помеченных данных), необходимой для построения решающего правила или обучения нейронных сетей, зачастую связан со значительными материальными и временными затратами. Поэтому на практике обучающая выборка (ОВ), как правило, имеется лишь для небольшого количества интересующих



а *б* *в*

Рисунок 5.7 – Исходное изображение (*а*) и результат его обработки алгоритмами MeanSC (*б*) и SPMeanSC (*в*)

пользователя классов и при этом является непредставительной (отдельные классы могут быть представлены несколькими помеченными пикселями).

Известно, что для обеспечения приемлемого качества классификации минимальное число точек ОВ для параметрических классификаторов составляет порядка $10k$ на класс (где k – размерность пространства признаков), а для непараметрических – $50k$ на класс. Поэтому проблема получения представительной ОВ особо актуальна при обработке гиперспектральных изображений, для которых число спектральных каналов (признаков) измеряется сотнями. В то же время при классификации изображений всегда доступен большой объём непомеченных данных. В этих условиях обучающую выборку можно расширить за счёт непомеченных данных с помощью методов, основанных на алгоритмах кластеризации.

Пакет «Image Processing Toolkit» включает алгоритм расширения обучающей выборки на основе результатов выполнения алгоритма SPMeanSC. Алгоритм SPMeanSCROIExt($h_r,\ h_s,\ \alpha,\ p$) можно условно разбить на три этапа.

На первом этапе изображение разбивается на сегменты с помощью алгоритма SPMeanSC с параметрами $\{h_r, h_s\}$.

На втором этапе для каждого класса исходной ОВ формируется множество сегментов, содержащих точки исходной ОВ для этого класса. Из полученных множеств удаляются сегменты, содержащие одновременно точки из нескольких классов исходной ОВ. Такие сегменты появляются при неудачном выборе параметров

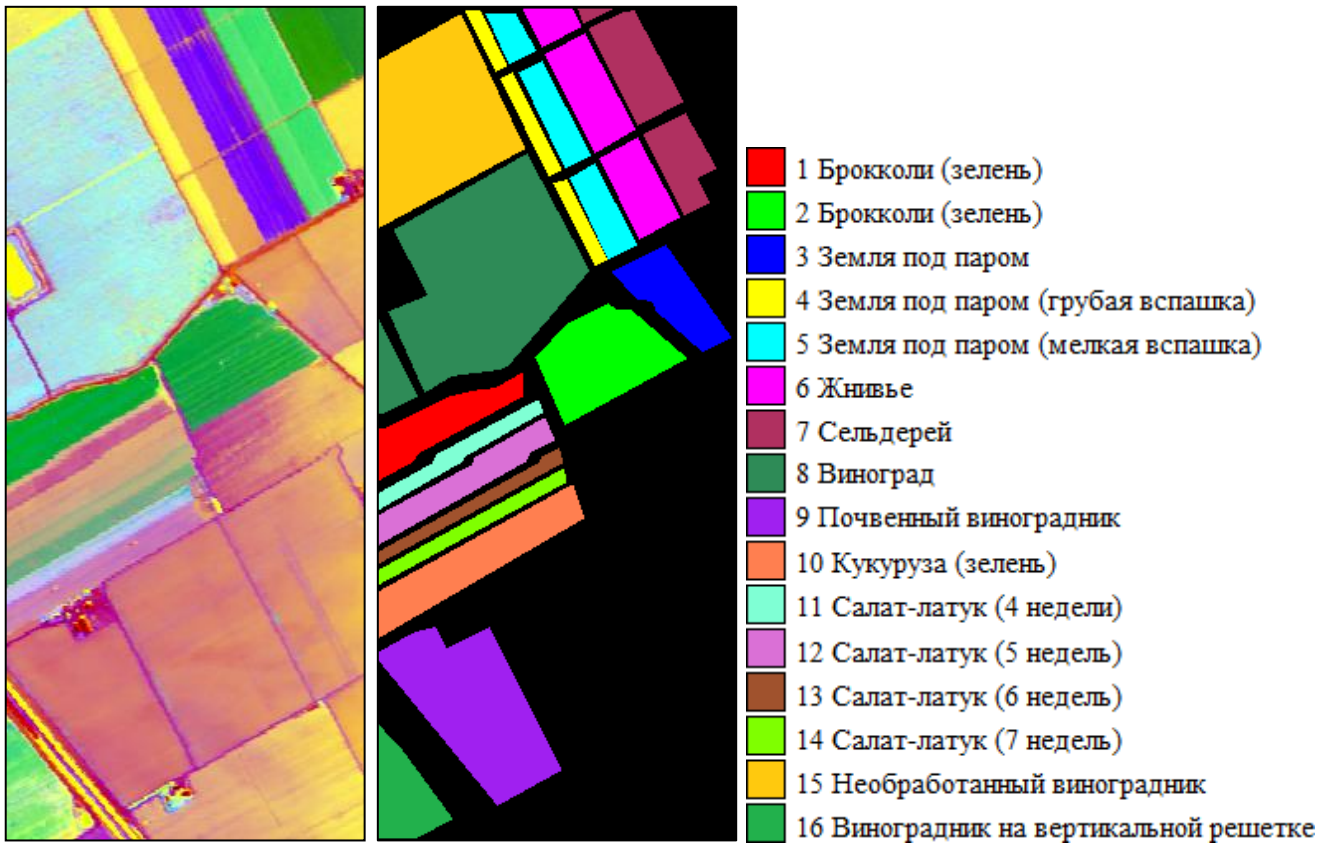
h_r и h_s, а также вследствие ошибок, допущенных при формировании исходной ОВ. Использование точек из таких сегментов для наращивания выборки может привести к грубым ошибкам классификации. Затем в каждое из сформированных множеств добавляются сегменты, расположенные в пространстве признаков на расстоянии не более, чем α от этого множества. Расстояние до множества вычисляется, как наименьшее из расстояний до его элементов; в качестве меры расстояния между сегментами используется евклидово расстояние между их представителями. После этого повторно удаляются сегменты, включённые одновременно в несколько множеств (расположенные в пространстве признаков на границе классов).

На последнем этапе алгоритма в обучающую выборку для каждого класса добавляются случайно выбранные точки из сегментов (p процентов от размера сегмента), которые включены в множество, сформированное для этого класса.

Для тестирования алгоритма использовалось известное гиперспектральное спутниковое изображение Salinas (долина Салинас, Калифорния) размером 512×217 пикселей, полученное с сенсора AVIRIS 8 октября 1998 года. Для обработки использовались признаки, построенные с помощью метода главных компонент. На рисунке 5.8 приведены RGB-композит изображения (первые три главных компоненты) и эталонная картосхема, включающая 16 классов.

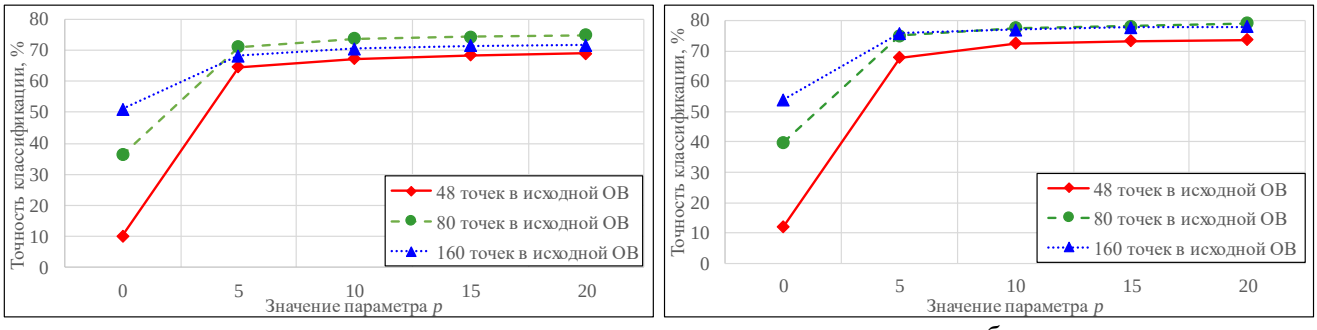
По эталонной картосхеме случайным образом были сформированы исходные обучающие выборки объёмом 48 точек (по 3 точки на каждый класс, представленный на эталонной картосхеме), 80 точек (по 5 точек на класс) и 160 точек (по 10 точек на класс). Каждая из этих выборок расширялась на 5, 10, 15 и 20 %. Значение параметра $\alpha = 10$ было зафиксировано. Расширение выборок осуществлялось с использованием первых четырёх признаков, полученных с помощью метода главных компонент. Переход к главным компонентам обусловлен высокой вычислительной сложностью процедуры «среднего сдвига», связанной с необходимостью многократного вычисления евклидова расстояния в многомерном пространстве признаков. Полученные выборки использовались для классификации изображения алгоритмом SVM (Support Vector Machine) на основе ра-

диальных базисных функций. Классификация выполнялась в пространстве признаков, полученных с помощью метода главных компонент. Использовалась реализация алгоритма SVM, включённая в пакет программ ITTVIS ENVI. Для всех настраиваемых параметров алгоритма выбирались значения по умолчанию. Точность классификации определялась посредством сравнения полученных картосхем



а *б*

Рисунок 5.8 – RGB-композит исходного изображения, составленный из первых трёх главных компонент (*а*), и эталонная картосхема, включающая 16 классов (*б*)



а *б*

Рисунок 5.9 – Средняя (*а*) и максимальная (*б*) точность классификации изображения Salinas алгоритмом SVM в зависимости от значения параметра p для разного объёма исходной ОВ. Значение $p = 0$ соответствует исходной выборке

с эталонной. Для усреднения результатов исходные выборки формировались по три раза и для каждой из этих выборок эксперимент с одним и тем же набором параметров повторялся трижды. Полученные значения точности классификации (в процентах) приведены в таблице. На рисунке 5.9 представлена зависимость средней и максимальной точности классификации от параметра p. Значение $p = 0$ соответствует исходной ОВ.

Анализ результатов показал, что расширение ОВ с помощью предложенного алгоритма позволяет значительно повысить точность классификации, особенно при скромном объёме исходной выборки (по 3 точки на класс). Кроме того, использование значений параметра $p > 10$ приводит к незначительному увеличению точности классификации при существенном росте времени обработки, особенно при большом объёме исходной выборки.

5.4. Решение практических задач

Разработанные алгоритмы и программное обеспечение прошло апробацию при решении двух практических задач.

1. Разделение формаций лесной растительности с близкими спектрально-яркостными характеристиками. На основе предложенных алгоритмов разработана методика автоматизированной обработки данных спутниковой съёмки, позволяющая успешно разделять формации растительности с близкими спектрально-яркостными характеристиками [151].

Предлагаемая процедура обработки изображения начинается с пороговой сегментации по нормализованному вегетационному индексу

$$\mathrm{NDVI} = \frac{I_{ir} - I_r}{I_{ir} + I_r},$$

где I_r, I_{ir} – значения сигналов, регистрируемые соответственно в красном и ближнем инфракрасном каналах спектра соответственно. Для более наглядного представления гистограммы, значения вегетационного индекса приводятся к интервалу

[0, 255]. Пороговые величины для отсечения непокрытых растительностью территорий (водная поверхность, минерализованные участки, постройки т.п.) определяются пользователем в диалоговом режиме на основе визуального анализа гистограммы значений NDVI. Затем производится линейное растяжение динамических диапазонов значений спектральных признаков на диапазон [0, 255].

После сегментации для всех возможных наборов признаков вычисляется значение OIF (Optimum Index Factor) [152]:

$$\mathrm{OIF}(S) = \frac{\sum_{i \in S} \sigma(i)}{\sum_{i \in S} \left(\sum_{j \in S, j > i} |r(i, j)| \right)},$$

где S – набор каналов; $\sigma(i)$ – среднеквадратичное отклонение значений спектральных яркостей для i-го канала; $r(i, j)$ – коэффициент корреляции значений спектральных яркостей между i-м и j-м каналами.

Затем определяются наборы признаков с наибольшими значениями OIF. Среди них, на основе анализа спектральных откликов растительности, определяется набор признаков, наиболее подходящий для решения конкретной задачи. Таким образом находятся слабо коррелированные и информативные признаки. Полученный набор признаков используется для классификации изображения при помощи непараметрического алгоритма кластеризации.

Методика прошла апробацию и доказала свою эффективность при исследовании лесной растительности Болотнинского района Новосибирской области. Изучаемый район ограничен 83˚49′53″ и 84˚29′7″ в.д. и 55˚58′42″ и 55˚52′42″ с.ш. Прямоугольниками выделены ключевые участки с координатами 83˚50′43″-83˚55′20″ в.д., 55˚54′45″-55˚53′15″ с.ш. и 84˚23′7″-84˚28′3″ в.д., 55˚59′23″-55˚57′26″ с.ш. Обработке подвергался фрагмент снимка Landsat 7 ETM+ с пространственным разрешением 30 м, полученного 1 августа 1999 г. (рисунок 5.10). Для дешифрирования фрагмента использовались данные наземных маршрутных наблюдений, собранные с использованием системы глобального позиционирования (GPS), а также почвенные карты Новосибирской области.

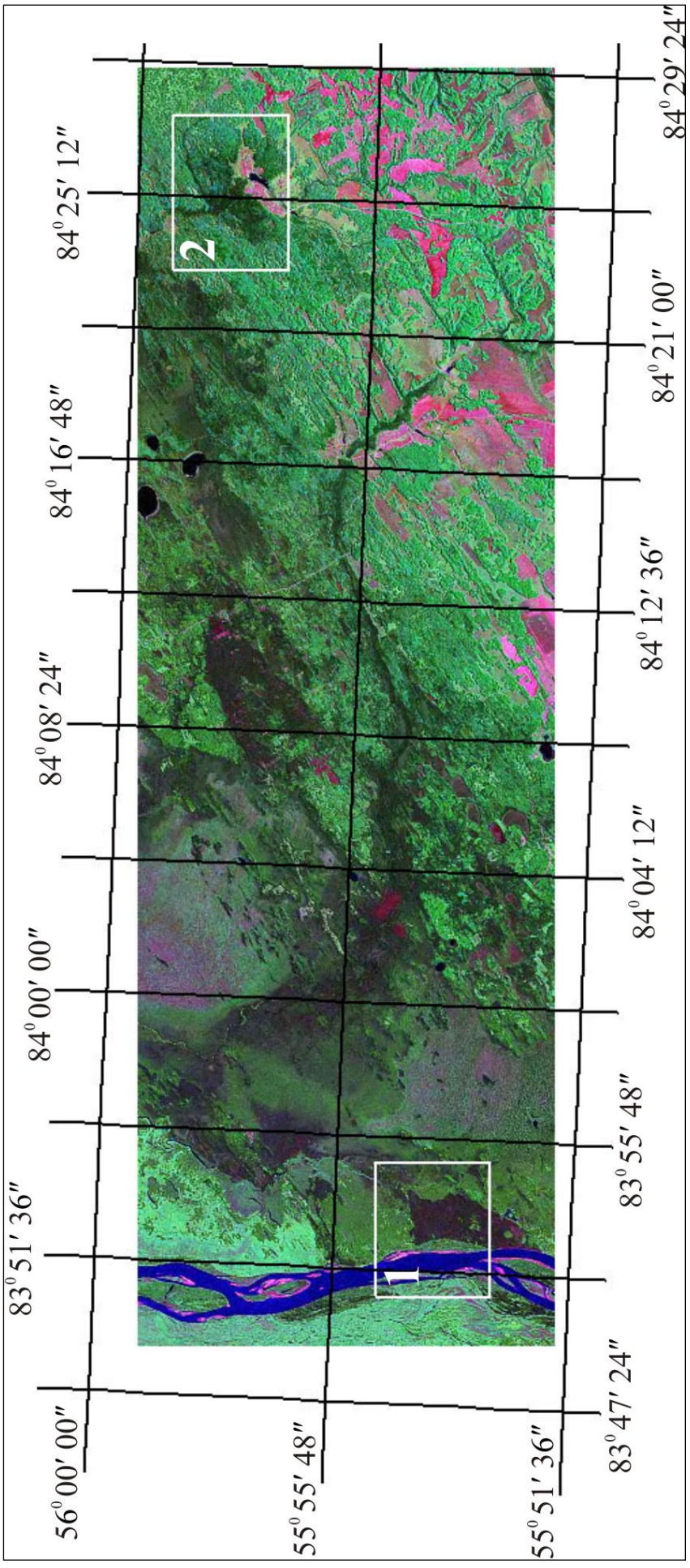
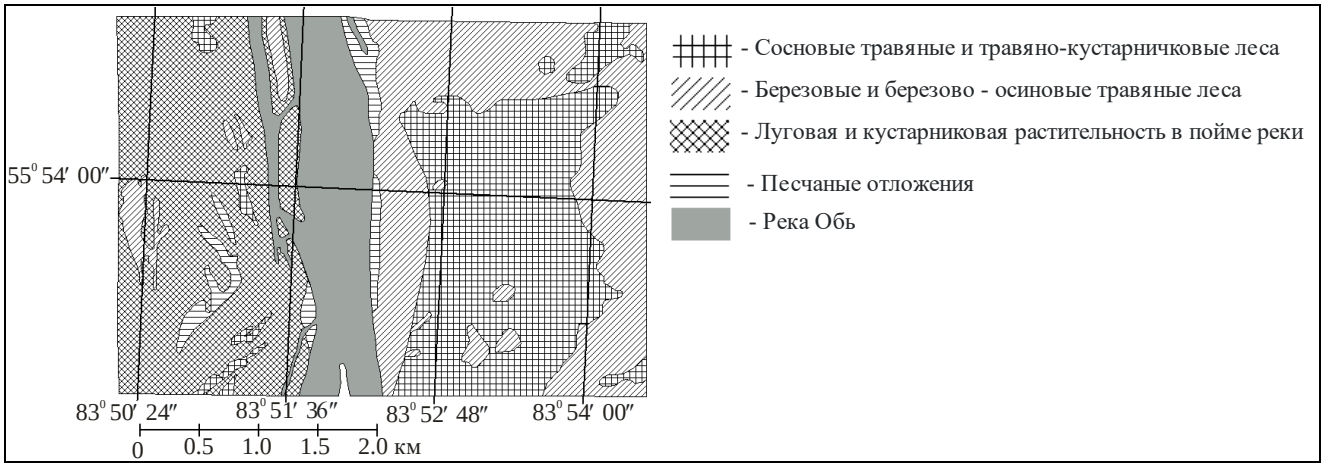
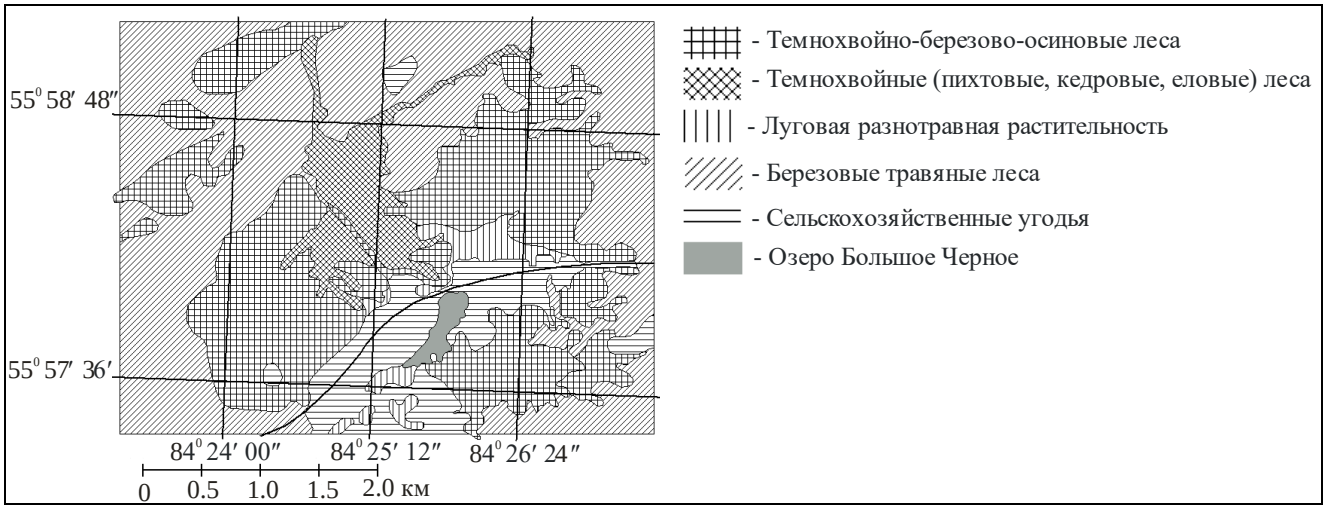


Рисунок 5.10 – Фрагмент изображения, полученного со спутника Landsat 7 ETM+, с нанесённой географической координатной сеткой. Прямоугольниками выделены ключевые участки, содержащие формации растительности с близкими спектрально-яркостными характеристиками



а



б

Рисунок 5.11 – Картосхемы участков 1 и 2 (*а* и *б* соответственно), полученные на основе визуально-инструментального дешифрирования снимка Landsat 7 ETM+

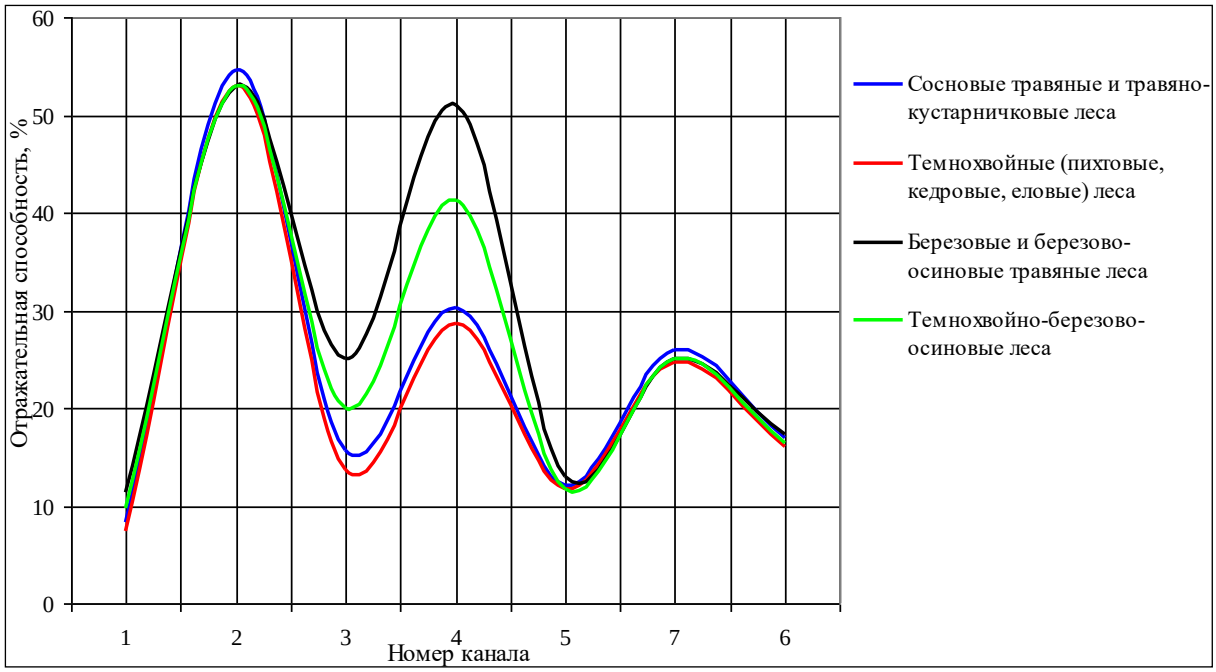


Рисунок 5.12 – Характеристики спектрального отклика лесной растительности

На рисунке 5.11 представлены картосхемы, полученные методом визуально-инструментального дешифрирования выделенных ключевых участков. С помощью этих картосхем по данным многозонального снимка были сформированы обучающие выборки для представленных типов лесной растительности во всех семи спектральных диапазонах. В результате анализа спектрально-яркостных характеристик лесной растительности, построенных по найденным обучающим выборкам (рисунок 5.12), установлено, что различные формации растительности имеют высокую спектральную селективность. Исключением являются темнохвойные (пихта, кедр, ель) и сосновые (травяные и травяно-кустарничковые) леса, имеющие близкие характеристики. Указанная особенность существенно усложняет разделение этих типов растительности без предварительной обработки снимка и выбора информативных признаков.

В соответствии с предложенной методикой была построена гистограмма значений вегетационного индекса (рисунок 5.13) и произведена сегментация изображения с пороговыми значениями 140 и 240 (рисунок 5.14). Это позволило снизить влияние неинтересующих территорий (не покрытых растительностью, на рисунке 5.14 изображены белым цветом) при определении информативных признаков с помощью величины OIF.

В таблице 5.1 представлены значения OIF для различных наборов признаков, вычисленные по сегментированному изображению. Анализ спектральных характеристик (см. рисунок 5.12) показывает, что для разделения темнохвойных и сосновых травяных лесов каналы 1 и 2 малоинформативны, а канал 3 значительно информативнее, чем канал 7. По этой причине среди наборов с наибольшим значением OIF для классификации выбран набор признаков {3, 4, 5}.

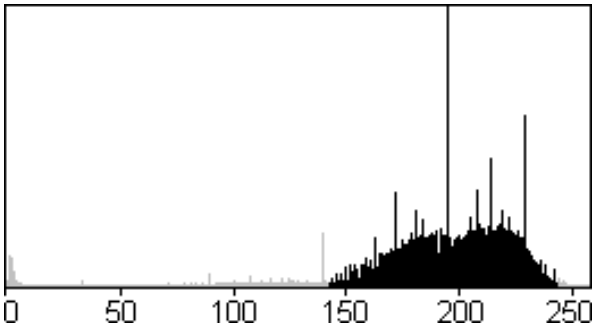


Рисунок 5.13 – Гистограмма

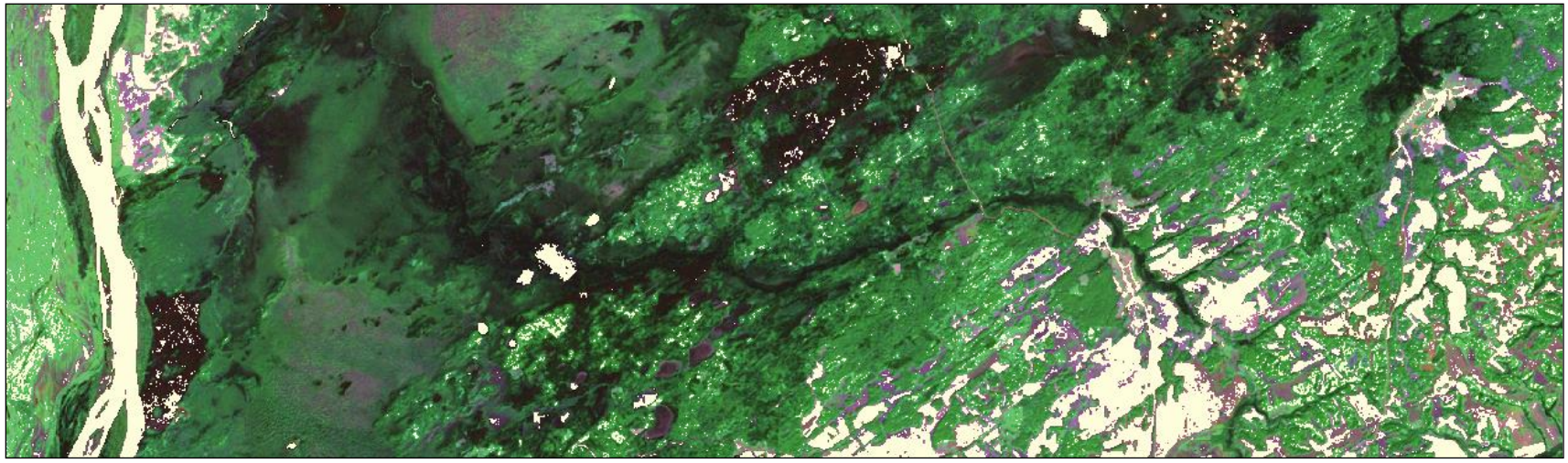


Рисунок 5.14 – Результаты пороговой сегментации по вегетационному индексу

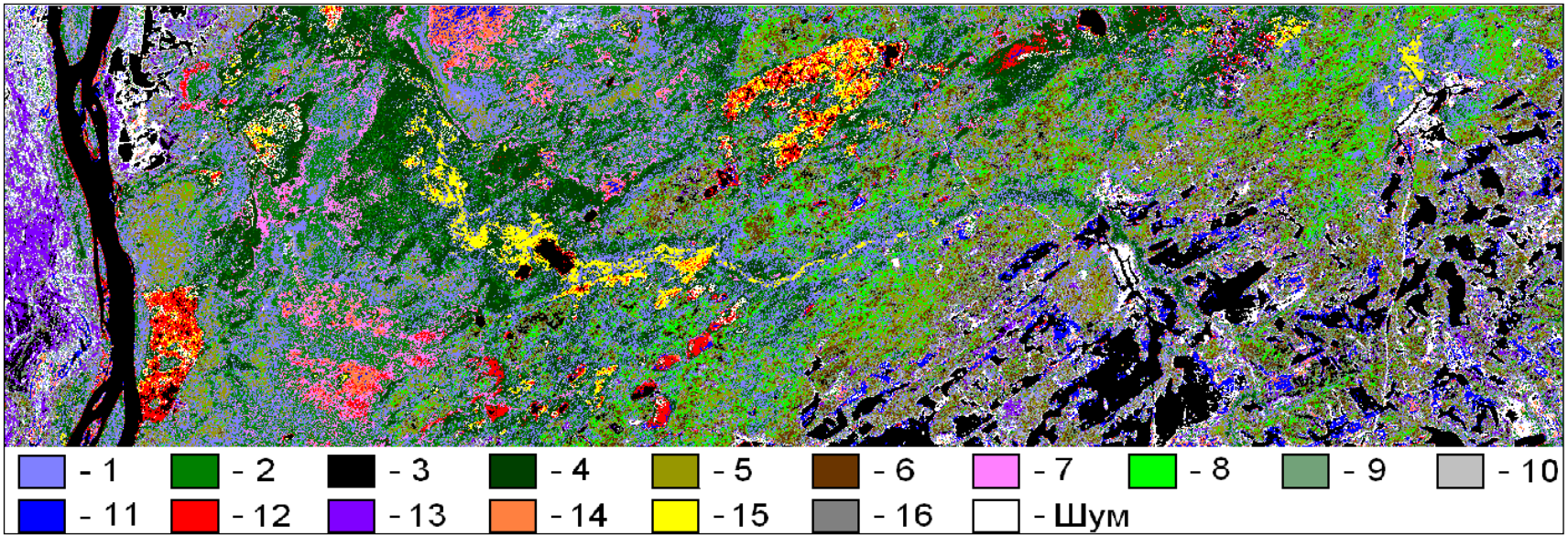


Рисунок 5.15 – Результаты классификации алгоритмом MeanSC

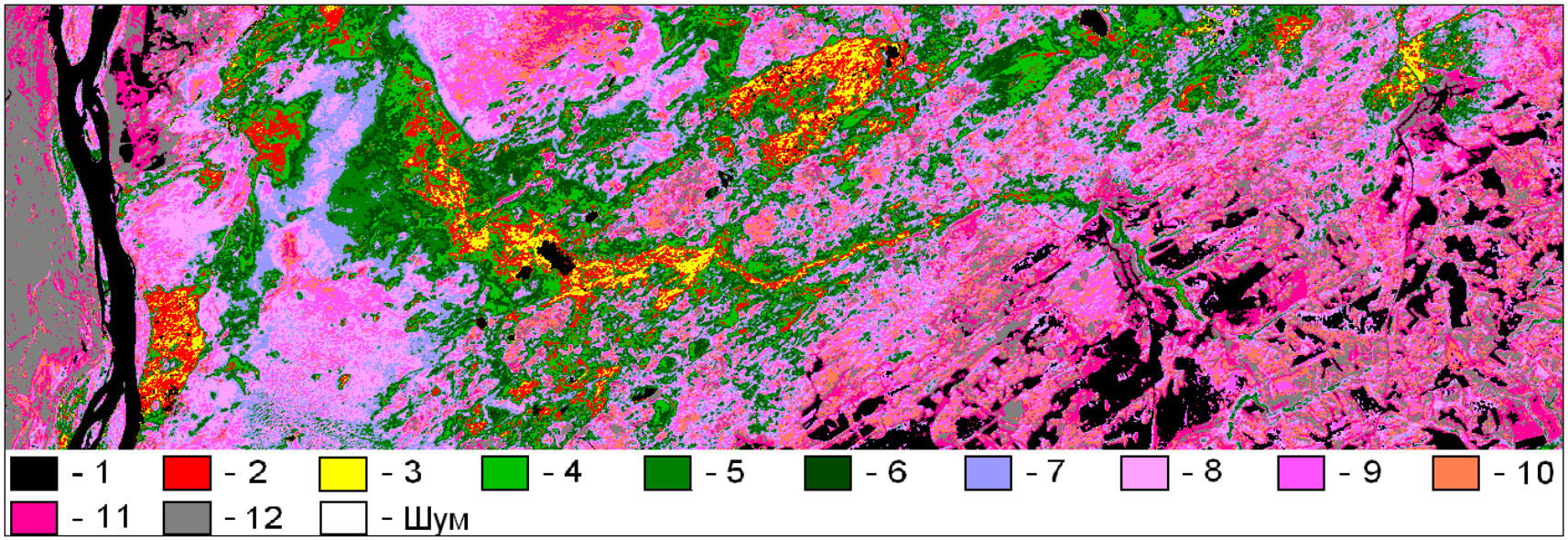


Рисунок 5.16 – Результаты классификации алгоритмом ISODATA

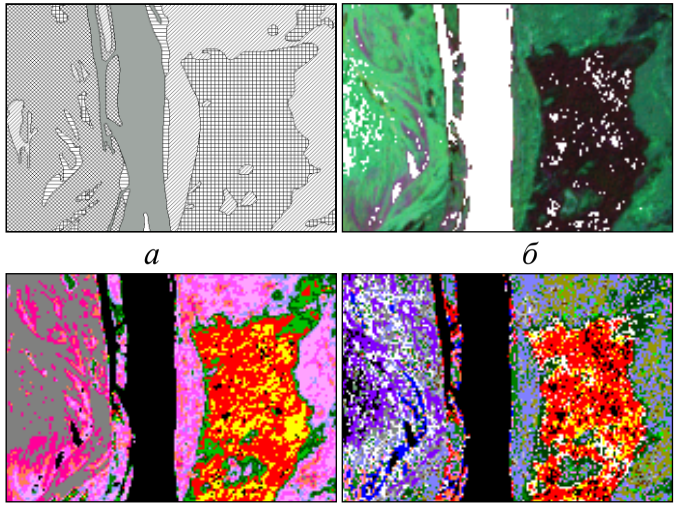


Рисунок 5.17 – Ключевой участок 1: исходные данные (*а*), результаты визуально-инструментального дешифрирования (*б*) и классификации с помощью алгоритмов MeanSC (*в*) и ISODATA (*г*)

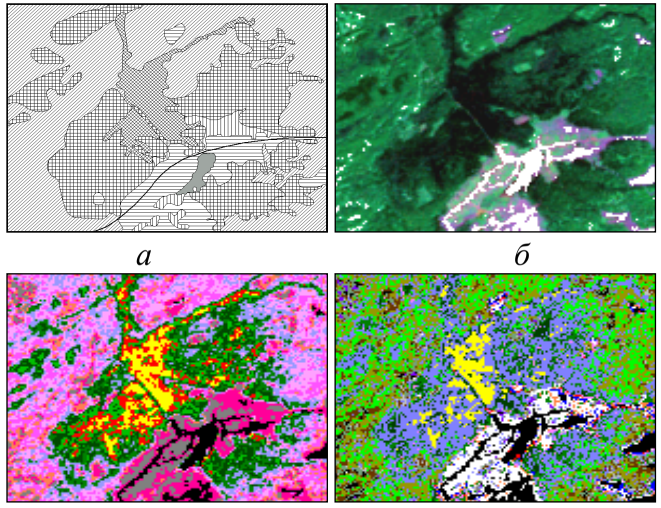


Рисунок 5.18 – Ключевой участок 2: исходные данные (*а*), результаты визуально-инструментального дешифрирования (*б*) и классификации с помощью алгоритмов MeanSC (*в*) и ISODATA (*г*)

Таблица 5.1 – Значения OIF для различных наборов из трёх признаков

Набор признаков	Величина OIF	Набор признаков	Величина OIF
{1, 4, 5}	29.7	{1, 2, 5}	26.6
{1, 4, 7}	27.1	{1, 5, 7}	19.5
{2, 4, 5}	26.8	{1, 3, 5}	18.7
{4, 5, 7}	26.2	{2, 5, 7}	16.7
{3, 4, 5}	26.1	{2, 3, 5}	16.2
{1, 2, 4}	25.4	{1, 2, 7}	16.2
{1, 3, 4}	25.1	{3, 5, 7}	15.2
{2, 4, 7}	24.1	{1, 3, 7}	14.9
{3, 4, 7}	23.0	{1, 2, 3}	14.7
{2, 3, 4}	22.7	{2, 3, 7}	12.6

По выбранным признакам проводилась обработка исследуемого фрагмента с помощью алгоритма MeanSC (рисунок 5.15) и алгоритма ISODATA, реализованного в пакете ENVI (рисунок 5.16). На рисунках 5.17 и 5.18 показано, что результаты кластеризации хорошо согласуются с результатами визуально-инструментального дешифрирования.

Проведенные исследования показали, что применение алгоритма ISODATA, а также алгоритма MeanSC непосредственно к изображению не позволяет разделить формации растительности, близкие по спектрально-яркостным характеристикам. Предварительная обработка изображения позволила улучшить результаты классификации с помощью алгоритма ISODATA. Однако и в этом случае не произошло полного разделения изучаемых формаций растительности. Предложенная методика на основе алгоритма MeanSC позволила разделить темнохвойные (пихта, кедр, ель) и сосновые (травяные и травяно-кустарничковые) леса.

2. Обнаружение усыхающих кедровых древостоев по мультиспектральным изображениям высокого пространственного разрешения. На основе ансамблевого алгоритма EMeanSC была разработана технология обнаружения и картирования повреждений кедровых древостоев по мультиспектральным изображениям высокого пространственного разрешения, полученным со спутника Pleiades [153].

Изучаемая зона расположена в горах Кузнецкого Алатау, на высотах от 640 до 1040 м над уровнем моря, и ограничена областью 54º29'24" и 54º31'12" с.ш., 88º48' и 88º52'12" в.д. (рисунок 5.19). Древостои сформированы кедром (>95%) с

примесью пихты и ели. Древостои спелые, средний возраст деревьев около 160 лет (максимальная продолжительность жизни кедра – 600–800 лет). Растительный покров мезофитный, с преобладанием осочки; толщина подстилки 3–5 см. Почвы светло-серого лесного типа глубиной 10–15 см, лежащие на слое каменистой глины. Признаков пожара внутри исследуемой зоны обнаружено не было. Наблюдения исследуемой зоны проводились с 2006 по 2012 годы. Исследованное усыхание кедровых древостоев восточного макросклона Кузнецкого Алатау является частью более широкого явления усыхания темнохвойных древостоев, сформированных кедром, пихтой и елью на территории Российской Федерации.

Исходным материалом для обработки послужило спутниковое изображение кедровых древостоев в бассейне р. Черный Июс (восточный макросклон Кузнецкого Алатау), полученный 1 июня 2012 года со спутника Pleiades (пространственное разрешение 2.8 м). Оценка точности картосхем, получаемых в результате обработки, выполнялась с использованием контрольных точек для трех классов интереса («усохший древостой», «вырубки усохшего древостоя на южном склоне» и «вырубки на северном склоне»), полученных в 2012 году в ходе проведения полевых исследований. Исходный снимок и схема расположения контрольных точек приведены на рисунке 5.20.

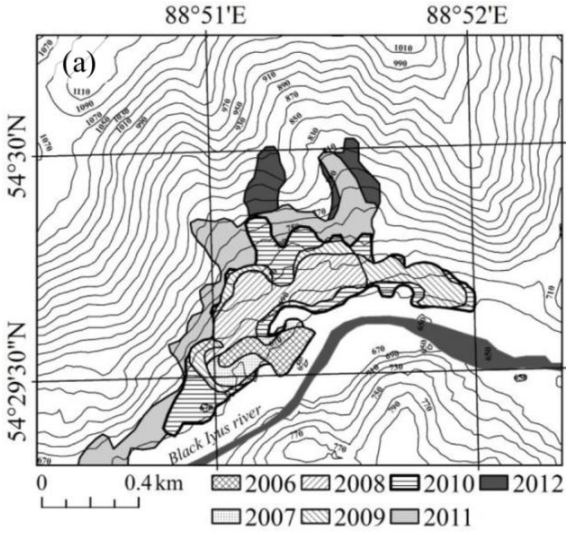


Рисунок 5.19 – Фрагмент картосхемы усыхания древостоев в исследуемой зоне с 2006 по 2012 годы

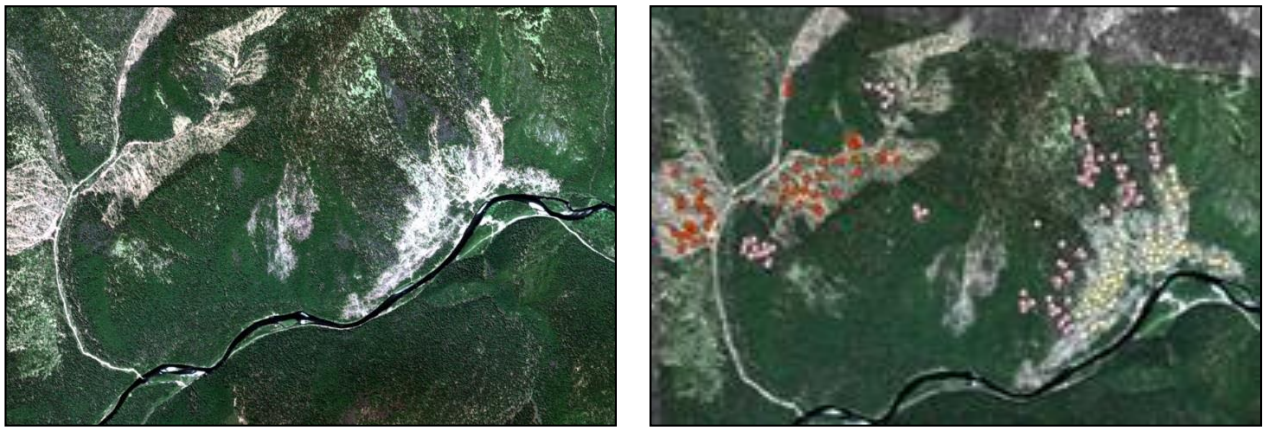


Рисунок 5.20 – Исходный снимок (слева) и схема расположения контрольных точек (справа), разбитых на три класса: «усохший древостой» (выделены розовым цветом), «вырубки усохшего древостоя на южном склоне» (выделены желтым цветом) и «вырубки на северном склоне» (выделены красным цветом)

Обработка изображения с целью обнаружения поврежденных территорий и их картирования состоит из трёх этапов. На первом этапе производится выбор информативной подсистемы признаков. Для этого к исходному набору спектральных каналов добавляется нормализованный вегетационный индекс (NDVI). Далее выполняется предварительная сегментация изображения быстрым сеточным алгоритмом автоматической классификации ECCA [116]. На построенной картосхеме выделяются однородные полигоны для формирования обучающих выборок. После этого для анализа информативности признаков и получения статистически «чистых» обучающих выборок используются гистограммы спектральных яркостей, построенные по выделенным однородным полигонам. В качестве критерия информативности выбрано расстояние Джеффриса – Матуситы [154], которое является оценкой верхней границы вероятности ошибки классификации. Второй этап заключается в попиксельной сегментации изображения по выбранным информативным признакам с помощью ансамблевого алгоритма кластеризации EMeanSC. Для уменьшения раздробленности и упрощения интерпретации итоговых картосхем на третьем этапе обработки применяется алгоритм генерализации и коррекции, учитывающий пространственный контекст (информацию о соседстве пикселей) [153].

В результате первого этапа обработки на изображении изучаемой зоны были выделены обучающие выборки для шести основных классов («усохший древостой», «вырубки усохшего древостоя на южном склоне», «вырубки на северном склоне», «здоровый древостой», «территории, не покрытые растительностью» и

«водная поверхность»). На рисунке 5.21 приведены гистограммы разброса значений спектральных яркостей для этих выборок. Анализ обучающих выборок показал, что классы «вырубки усохшего древостоя на южном склоне» и «вырубки на северном склоне» очень близки и их следует объединить в один класс интереса – «вырубки». Пример однородных полигонов и полученной обучающей выборки для класса «территория без растительности» приведен на рисунке 5.22.

Среднее значение расстояния Джеффриса – Матуситы J между классами при использовании пяти признаков составило 1.8. Четыре признака (1, 3 и 4 каналы и NDVI) позволяют достичь $J = 1.78$. Комбинации из трех признаков также достаточно информативны ($J = 1.71$), хотя минимальное расстояние между классами уменьшается. Дальнейшее уменьшение размерности системы признаков ведет к снижению информативности ($J = 1.49$ для двух и $J = 0.5$ для одного признака). В итоге, оптимальной была выбрана подсистема из четырех признаков: 1, 3, 4 каналы и NDVI. Заметим, что 4 канал и NDVI часто присутствовали в комбинациях признаков с высоким значением критерия информативности. Учет контекстной информации значительно облегчил процесс интерпретации картосхем, полученных на втором этапе обработки (рисунок 5.23).

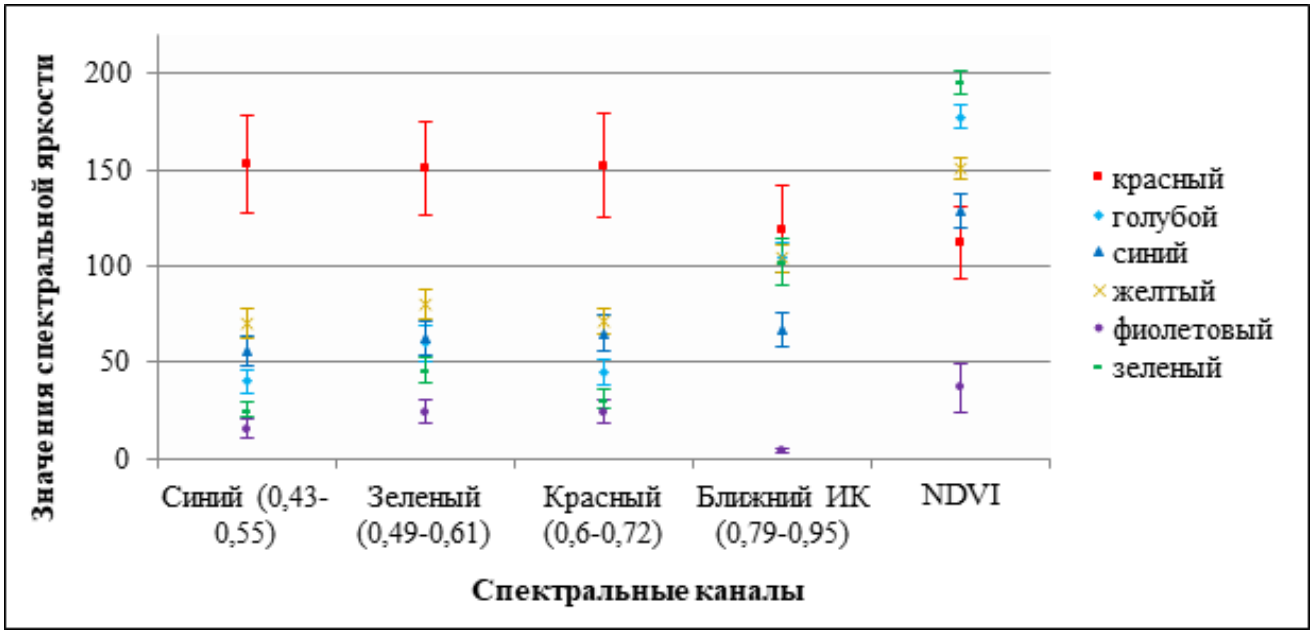


Рисунок 5.21 – Диаграмма разброса значений спектральных яркостей для классов: красный – «территории, не покрытые растительностью»; голубой – «вырубки усохшего древостоя на южном склоне»; синий – «усохший древостой»; желтый – «вырубки на северном склоне»; фиолетовый – «водная поверхность»; зеленый – «здоровый древостой»

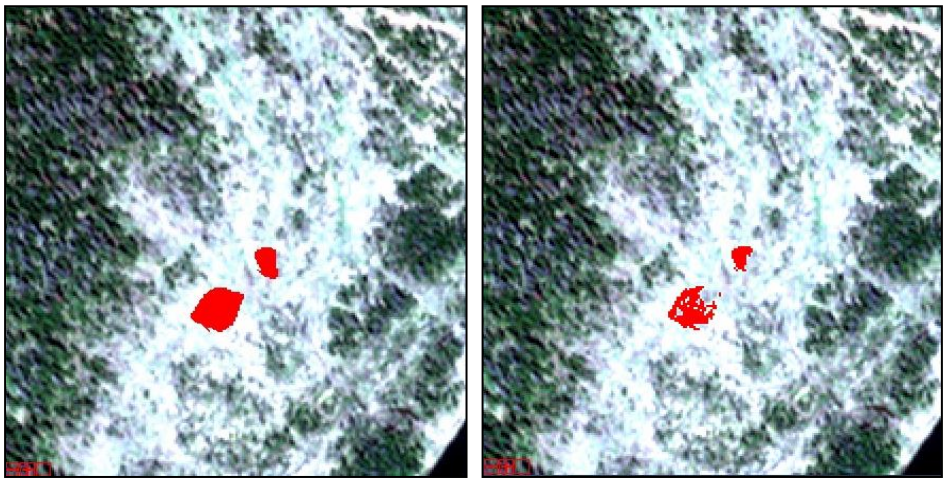
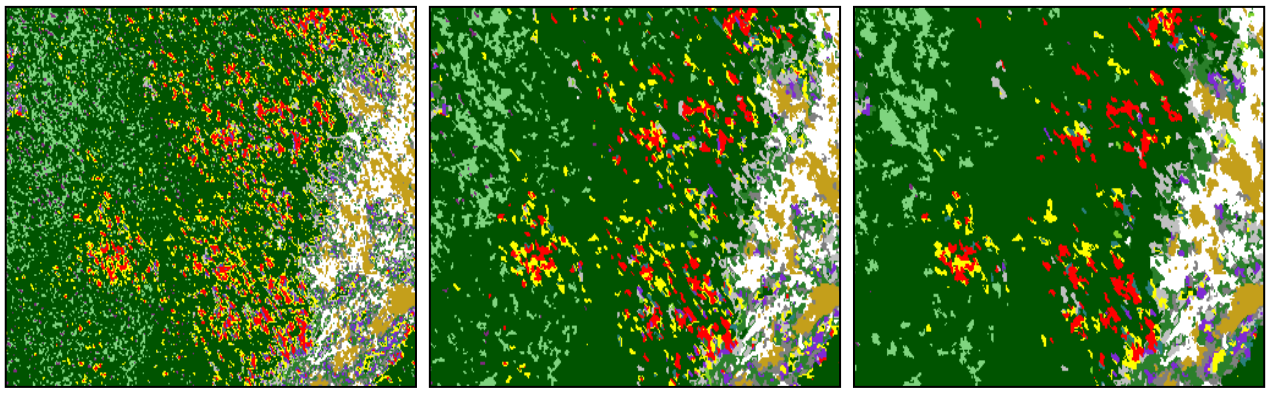


Рисунок 5.22 – Однородные полигоны (слева) и обучающая выборка (справа) для класса «территории, не покрытые растительностью»



а *б* *в*

Рисунок 5.23 – Обработка фрагмента картосхемы, полученной с помощью алгоритма EMeanSC (*а*), контекстным алгоритмом кластеризации с использованием в качестве маркеров 10% (*б*) и 5% (*в*) пикселов изображения

Таблица 5.2 – Распределение данных полевых наблюдений по кластерам

Номер кластера	1	2	3	4	5	6	7
Число точек из класса «вырубки»	7	82	13	6	0	0	4
Число точек из класса «усохший лес»	0	0	0	13	70	15	0

С использованием результатов полевых наблюдений (268 тестовых точек, разбитых на 2 класса интереса – «вырубки» и «усохший лес»), проведена оценка достоверности сегментации (таблица 5.2). Класс интереса «вырубки» был разбит алгоритмом EMeanSC на четыре кластера (кластеры 1-3 и 7) а класс «усохший лес» – на три (кластеры 4-6). При этом 6 точек из класса «усохший лес» (кластер 4) были ошибочно отнесены к классу «вырубки».

Данная методика обработки позволила с высокой степенью достоверности выявить очаги повреждений кедровых древостоев на восточном макросклоне Кузнецкого Алатау (бассейн реки Чёрный Июс) по снимку со спутника Pleiades.

Представленные результаты апробации на практических задачах демонстрируют эффективность разработанного программно-алгоритмического инструментария в задачах, связанных с обнаружением и выделением мелких и трудноразделимых классов на изображении, что очень важно при построении детальных картографических моделей местности и обнаружении повреждений растительности на ранней стадии.

Выводы по главе

1. На основе программных продуктов с открытым исходным кодом создана платформа для публикации алгоритмов обработки пространственных данных в виде стандартизованных веб-сервисов (WPS-процессов). В виде WPS-процессов опубликовано пять алгоритмов классификации и кластеризации, в том числе MeanSC и EMeanSC. Платформа доступна по адресу http://wps.ict.nsc.ru:8080/wps/WebProcessingService (протокол доступа – WPS).
2. Разработан модуль для геоинформационной системы с открытым исходным кодом GRASS GIS, реализующий алгоритм EMeanSC.
3. Создан пакет программ для обработки и анализа мультиспектральных изображений «Image Processing Toolkit», который включает алгоритмы MeanSC и EMeanSC, а также непараметрический алгоритм сегментации изображений SPMeanSC, учитывающий пространственную информацию, и алгоритм наращивания обучающей выборки, разработанные автором.
4. На основе алгоритмов MeanSC и EMeanSC разработаны методы разделения формаций лесной растительности с близкими спектрально-яркостными характеристиками и обнаружения усыхающих древостоев по мультиспектральным изображениям. Эти методы позволяют обеспечить качественное выделение мелких и сильно пересекающихся классов, которые не обнаруживаются при использовании традиционных методов автоматизированный обработки.

ЗАКЛЮЧЕНИЕ

Данная диссертационная работа была направлена на разработка эффективных непараметрических алгоритмов сегментации спутниковых изображений и современной платформы для стандартизованного доступа к ним, обеспечивающей разработчикам простой и быстрый механизм внедрения новых алгоритмов, а потенциальным пользователям – прозрачный доступ к опубликованным методам. Итогом работы стали следующие научные и практические результаты.

1. Разработан и исследован вычислительно эффективный непараметрический алгоритм кластеризации MeanSC на основе оценок плотности Розенблатта – Парзена для сегментации мультиспектральных спутниковых изображений. Эффективность достигается благодаря введению сеточной структуры в пространстве признаков и переходу к рабочей выборке значительно меньшего объёма, в которой гарантированно содержатся представители всех классов, присутствующих на изображении. Сеточная структура в пространстве признаков впервые использована для повышения вычислительной эффективности поэлементного алгоритма кластеризации.
2. Предложен подход к построению ансамбля непараметрических алгоритмов кластеризации, основанных на оценках плотности Розенблатта – Парзена, с помощью согласованной матрицы различий. В рамках этого подхода на основе алгоритма MeanSC разработан ансамблевый алгоритм кластеризации EMeanSC, позволяющий обеспечить простоту настройки параметров и обработку мультиспектральных спутниковых изображений в диалоговом режиме.
3. На основе предложенных алгоритмов кластеризации разработаны методы разделения формаций лесной растительности с близкими спектрально-яркостными характеристиками и обнаружения усыхающих древостоев по мультиспектральным изображениям. Эти методы позволяют обеспечить качественное выделение мелких и сильно пересекающихся классов, которые не обнаруживаются при использовании традиционных методов автоматизированной обработки.

4. Выполнен сравнительный анализ предложенных алгоритмов MeanSC и EMeanSC с алгоритмами, включёнными в широко распространённый пакет для обработки спутниковых данных ENVI и в пакеты для анализа данных ELKI и Smile. На модельных данных показано, что алгоритмы MeanSC и EMeanSC превосходят известные непараметрические алгоритмы по точности и/или быстродействию. На реальных изображениях продемонстрировано, что разработанные непараметрические алгоритмы позволяют обрабатывать мультиспектральные изображения в диалоговом режиме.
5. На основе программных продуктов с открытым исходным кодом создана платформа для публикации алгоритмов обработки пространственных данных в виде стандартизованных веб-сервисов (WPS-процессов). В виде WPS-процессов опубликовано пять алгоритмов классификации и кластеризации, в том числе MeanSC и EMeanSC. Платформа доступна по адресу http://wps.ict.nsc.ru:8080/wps/WebProcessingService (протокол доступа – WPS). Разработан модуль для геоинформационной системы с открытым исходным кодом GRASS GIS, реализующий алгоритм EMeanSC.
6. Создан пакет программ для обработки и анализа мультиспектральных изображений «Image Processing Toolkit», который включает алгоритмы MeanSC и EMeanSC, а также непараметрический алгоритм сегментации изображений SPMeanSC, учитывающий пространственную информацию, и алгоритм наращивания обучающей выборки, разработанные автором. Пакет «Image Processing Toolkit» передан в Институт почвоведения и агрохимии СО РАН, где используется при крупномасштабном картографическом моделировании структурной организации растительности и почвенного покрова. Результаты его апробации показали эффективность разработанного программно-алгоритмического инструментария при решении практических задач, связанных с анализом мультиспектральных спутниковых изображений.

СПИСОК ЛИТЕРАТУРЫ

1. Dey, V. A review on image segmentation techniques with remote sensing perspective / V. Dey, Y. Zhang, M. Zhong // Proceedings of the ISPRS TC VII Symposium «100 Years ISPRS». – Vienna, Austria, July 5-7, 2010. – Vol. XXXVIII. – Part 7A. – P. 31-42.
2. Rekik, A. Review of satellite image segmentation for an optimal fusion system based on the edge and region approaches / A. Rekik, M. Zribi, A. Hamida, M. Benjellounl // IJCSNS International Journal of Computer Science and Network 242 Security. – 2007. – Vol. 7. – No. 10. – P. 242-250.
3. Применение данных ДЗЗ [Электронный ресурс]. – 2011. – Режим доступа: http://www.sovzond.ru/about/publications/543.
4. Гонсалес, Р. Цифровая обработка изображений / Р. Гонсалес, Р. Вудс. – Москва: Техносфера, 2006. – 1104 с.
5. Шапиро, Л. Компьютерное зрение / Л. Шапиро, Дж. Стокман. – Москва: БИНОМ, Лаборатория знаний, 2006. – 752 с.
6. Sarmah, S. A grid-density based technique for finding clusters in satellite image / S. Sarmah, D. K. Bhattacharyya // Pattern Recognition Letters. – 2012. – Vol. 33. – P. 589-604.
7. Кудашев, Е. Б. Развитие инфраструктуры распределенных хранилищ спутниковых данных: интегрированная распределенная среда неод-нородных информационных ресурсов исследования Земли из Космоса / Е. Б. Кудашев, А. Н. Филонов // Труды Десятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2008). – Дубна, Россия, 2008. – С. 299-308.
8. Шокин, Ю. И. Распределённая информационная система сбора, хранения и обработки спутниковых данных для мониторинга территорий Сибири и Дальнего Востока / Ю. И. Шокин, И. А. Пестунов, В. В. Смирнов и др. // Журнал Сибирского федерального университета. Техника и технологии. – 2008. – Т. 1. – Выпуск 4. – С. 291-314.

9. Web feature service [Electronic resource]. – 2012. – URL: http://www.opengeospatial.org/standards/wfs.
10. Web map service [Electronic resource]. – 2012. – URL: http://www.opengeospatial.org/standards/wms.
11. Web processing service [Electronic resource]. – 2012. – URL: http://www.opengeospatial.org/standards/wps.
12. OGC standards and specifications [Electronic resource]. – 2012. – URL: http://www.opengeospatial.org/standards
13. Жижимов, О. Л. Принципы построения распределенных информационных систем на основе протокола Z39.50 / О. Л. Жижимов, Н. А. Мазов. – Новосибирск: ОИГГМ СО РАН; ИВТ СО РАН, 2004. – 361 с.
14. Смирнов, В. В. Корпоративные картографические сервисы Сибирского отделения РАН / В. В. Смирнов, И. А. Пестунов, Д. И. Добротворский, Ю. Н. Синявский // Горный информационно-аналитический бюллетень. – 2009. – Выпуск 18 «Кузбасс 3». – С. 130-134.
15. Jain, A. K. Data clustering: 50 years beyond k-means / A. K. Jain // Pattern Recognition Letters. – 2010. – Vol. 31. – No. 8. – P. 651-666.
16. Jain, A. K. Dimensionality and sample size considerations in pattern recognition practice / A. K. Jain, B. Chandrasekaran; P. R. Krishnaiah, L. N. Kanal, eds. – Handbook of Statistics. – Amsterdam: North-Holland, 1982. – Vol. 2. – P. 835-855.
17. Jain, A. K. Feature selection: evaluation, application, and small sample performance / A. K. Jain, D. Zongker // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1997. – Vol. 19. – No. 2. – P. 153-158.
18. Friedman, J. H. Exploratory projection pursuit / J. H. Friedman // Journal of the American Statistical Association. – 1987. – Vol. 82. – P. 249-266.
19. Тарасенко, Ф. П. Непараметрическая статистика / Ф. П. Тарасенко. – Томск: Издательство ТГУ, 1976. – 294 с.
20. Адаптивные системы и их приложения / под ред. А. В. Медведева. – Новосибирск: Наука, 1978. – 191 с.

21. Ghosh, S. Understanding deep learning techniques for image segmentation / S. Ghosh, N. Das, I. Das, U. Maulik // ACM Computing Surveys. – 2019. – Vol. 52. – No. 4. – P. 73:1-73:35.
22. Xu, D. A comprehensive survey of clustering algorithms / D. Xu, Y. Tian // Annals of Data Science. – 2015. – Vol. 2. – P. 165-193.
23. Nerurkara, P. Empirical analysis of data clustering algorithms / P. Nerurkara, A. Shirkeb, M. Chandanec, S. Bhirud // Procedia Computer Science. – 2018. – Vol. 125. – P. 770-779.
24. Cohen-Addad, V. Hierarchical clustering: objective functions and algorithms [Electronic resource] / V. Cohen-Addad, V. Kanade, F. Mallmann-Trenn, C. Mathieu // Journal of the ACM (JACM). – 2019. – Vol. 66. – No. 4. – Article No. 26. – 42 p.
25. Chouhan, S. S. Image segmentation using computational intelligence techniques: review / S. S. Chouhan, A. Kaul, U. P. Singh // Archives of Computational Methods in Engineering. – 2019. – Vol. 26. – No. 3. – P. 533-596.
26. Liu, X. Recent progress in semantic image segmentation / X. Liu, Z. Deng, Y. Yang // Artificial Intelligence Review. – 2019. – Vol. 52. – P. 1089-1106.
27. Ahmed, N. Recent review on image clustering / N. Ahmed // IET Image Processing. – 2015. – Vol. 9. – No. 11. – P. 1020-1032.
28. Bouguettaya, A. A comparison of group-based and object-based data clustering techniques / A. Bouguettaya, Q. Le Viet, M. Golea // Eighth International Database Workshop, Data Mining, Data Warehousing and Client/Server Databases. – Hong Kong: Springer-Verlag Singapore, 1997. – P. 119-136.
29. Ilango, Mr. A survey of grid based clustering algorithms / Mr. Ilango, V. Mohan // International Journal of Engineering Science and Technology. – 2010. – Vol. 2(8). – P. 3441-3446.
30. Jain, A. K. Statistical pattern recognition: A review / A. K. Jain, R. P. W. Duin, J. Mao // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2000. – Vol. 22. – No. 1. – P. 4-37.
31. Jain, A. K. Data clustering: A review / A. K. Jain, M. N. Murty // ACM Computing Surveys. – 1999. – Vol. 31. – No. 3. – P. 264-323.

32. Mercer, D. P. Clustering large datasets [Electronic resource] / D. P. Mercer. – Linacre College, 2003. – URL: http://ldc.usb.ve/~mcuriel/Cursos/WC/Transfer.pdf.
33. Parsons, L. Evaluating subspace clustering algorithms / L. Parsons, E. Haque, H. Liu // Workshop on Clustering High Dimensional Data and its Applications. – SIAM International Conference on Data Mining (SDM-2004). – 2004. – P. 48-56.
34. Parsons, L. Subspace clustering for high dimensional data: A review / L. Parsons, E. Haque, H. Liu // SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining. – 2004. – Vol. 6. – No. 1. – P. 90-105.
35. Xu, R. Survey on clustering algorithms / R. Xu, D. C. II. Wunsch // IEEE Trans. On Neural Networks. – 2005. – Vol. 16. – No. 3. – P. 645-678.
36. Luo, L. Nonparametric Bayesian correlated group regression with applications to image classification / L. Luo, J. Yang, B. Zhang, J. Jiang, H. Huang // IEEE Transactions on Neural Networks and Learning Systems. – 2018. – Vol. 29. – No. 11. – P. 5330-5344.
37. Wang, Y.-X. Noisy sparse subspace clustering / Y.-X. Wang, H. Xu // Journal of Machine Learning Research. – 2016. – Vol. 17. – P. 1-41.
38. Ezugwu, A. E. Nature-inspired metaheuristic techniques for automatic clustering: A survey and performance study [Electronic resource] / A. E. Ezugwu // SN Applied Sciences. – 2020. – Vol. 2. – ArticleID 273. – URL: https://link.springer.com/content/pdf/10.1007%2Fs42452-020-2073-0.pdf.
39. Aggarwal, C. C. Data clustering: Algorithms and applications / C. C. Aggarwal, C. K. Reddy. – Chapman and Hall, 2014. – 648 p.
40. Anderberg, M. R. Cluster analysis for applications / M. R. Anderberg. – Academic press, 1973.
41. Gan, G. Data clustering: Theory, algorithms, and applications / G. Gan, C. Ma, J. Wu. – ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007. – 466 p.
42. Hartigan, J. A. Clustering algorithms / J. A. Hartigan. – N.Y.: John Wiley & Sons, 1975. – 351 p.

43. Xu, R. Clustering / R. Xu, D. C. II. Wunch. – N.Y.: John Wiley & Sons, 2009. – 358 p.
44. Дидэ, Э. Методы анализа данных: Подход, основанный на методе динамических сгущений / Э. Дидэ; С. А. Айвазян, В. М. Бухштабер (ред.). – Москва: Финансы и статистика, 1985. – 357 с.
45. Дюран, Н. Кластерный анализ / Н. Дюран, П. Оделл. – М.: Статистика, 1977. – 128 с.
46. Миркин, Б. Г. Группировки в социально-экономических исследованиях: Методы построения и анализа / Б. Г. Миркин. – М.: Финансы и статистика, 1985. – 223 с.
47. Duda, R. Pattern classification. 2nd ed. / R. Duda, P. Hart, D. Stork. – N.Y.: John Wiley & Sons, 2001. – 688 p.
48. Айвазян, С. А. Прикладная статистика: классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. – Москва: Финансы и статистика, 1989. – 607 с.
49. Загоруйко, Н. Г. Прикладные методы анализа данных и знаний / Н. Г. Загоруйко. – Новосибирск: Издательство Института математики, 1999. – 270 с.
50. Ту, Дж. Принципы распознавания образов / Дж. Ту, Р. Гонсалес. – Москва: Мир, 1978. – 411 с.
51. Ankerst, M. OPTICS: ordering points to identify the clustering structure / M. Ankerst, M. M. Breunig, H.-P. Kriegel, J. Sander // Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. – ACM Press, 1999. – P. 49-60.
52. Brecheisen, S. Density-based data analysis and similarity search / S. Brecheisen, H.-P. Kriegel, P. Kroger et al. // In: Petrushin V. A., Khan L. (eds.): Multimedia Data Mining and Knowledge Discovery. – Springer, 2006. – P. 94-115.
53. Du, K.-L. Clustering: A neural network approach / K.-L. Du // Neural Networks. – 2010. – Vol. 23. – P. 89-107.

54. Cutting, D. Scatter/gather: A cluster-based approach to browsing large document collections / D. Cutting, D. Karger, J. Pedersen, J. Tukey // Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – Copenhagen, Denmark, 1992. – P. 318-329.
55. Tantrum, J. Model-based clustering of large datasets through fractionization and refractionization / J. Tantrum, A. Murua, W. Stuetzle // Proceedings of the ACM SIGKDD Conference. – Edmonton, Alberta, Canada, 2002. – P. 183-190.
56. Zhang, T. BIRCH: An efficient data clustering method for very large databases / T. Zhang, R. Ramakhrisnan, M. Livny // Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'96). – 1996. – P. 103-114.
57. Guha, S. CURE: An Efficient Clustering Algorithm for Large Databases / S. Guha, R. Rastogi, K. Shim // Proceedings of the ACM SIGMOD International Conference on Management of Data. – ACM Press, 1998. – P. 73-84.
58. Efimov, K. Adaptive nonparametric clustering / K. Efimov, L. Adamyan, V. Spokoiny // IEEE Transactions on Information Theory. – 2019. – Vol. 65. – No. 8. – P. 4875-4892.
59. Forgy, E. Cluster analysis of multivariate data: efficiency vs. interpretability of classifications / E. Forgy // Biometrics. – 1965. – Vol. 21. – P. 768-780.
60. Selim, S. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality / S. Selim, M. Ismail // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1984. – Vol. 6. – No. 1. – P. 81-87.
61. Pal, P. A symmetry based clustering technique for multi-spectral satellite imagery [Electronic resource] / P. Pal, B. Chanda // Proceedings of the Third Indian Conference on Computer Vision, Graphics and Image Processing. – 2002. – URL: http://www.ee.iitb.ac.in/~icvgip/PAPERS/252.pdf.
62. Ball, G. A clustering technique for summarizing multivariate data / G. Ball, D. Hall // Behavioral Science. – 1967. – Vol. 12. – P. 153-155.
63. Ёлкин, Е. А. О возможности применения методов распознавания в палеонтологии / Е. А. Ёлкин, В. Н. Ёлкина, Н. Г. Загоруйко // Геология и геофизика. – 1967. – № 9. – С. 75-78.

64. Kauffmanб L. Finding groups in data: an introduction to cluster analysis / L. Kauffman, P. J. Rousseuw. – John Wiley & Sons, 2005. – 342 p.
65. Ng, R. T. Efficient and effective clustering methods for spatial data mining / R. T. Ng, J. Han // Proceedings of the 20th VLDB Conference – 1994. – P. 144-155.
66. Aggarwal, C. C. Fast algorithms for projected clustering / C. C. Aggarwal, J. L. Wolf, P. S. Yu et al. // Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. – ACM Press, 1999. – P. 61-72.
67. Aggarwal, C. C. Finding generalized projected clusters in high dimensional spaces / C. C. Aggarwal, P. S. Yu // Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. – ACM Press, 2000. – P. 70-81.
68. Woo, K.-G. FINDIT: A fast and intelligent subspace clustering algorithm using dimension voting / K.-G. Woo, J.-H. Lee, M.-H. Kim, Y.-J. Lee // Information and Software Technology. – 2004. – Vol. 46. – No. 4. – P. 255-271.
69. Leopolda, N. UNIC: a fast nonparametric clustering [Electronic resource] / N. Leopolda, O. Rose // Pattern Recognition. – 2020. – Vol. 100. – URL: https://www.sciencedirect.com/science/article/pii/S0031320319304182.
70. Titterington, D. Statistical analysis of finite mixture distributions / D. Titterington, A. Smith, U. Makov. – Chichester, U.K.: John Wiley & Sons, 1985. – 243 p.
71. Деврой, Л. Непараметрическое оценивание плотности. L_1-подход. / Л. Деврой, Л. Дьёрфи. – М.: Мир, 1988. – 408 с.
72. Narendra, P. M. A non-parametric clustering scheme for LANDSAT / P. M. Narendra, M. Goldberg // Pattern Recognition. – 1977. – Vol. 9. – P. 207.
73. Сидорова, В. С. Анализ многоспектральных данных дистанционного зондирования покрова Земли с помощью гистограммного иерархического кластерного алгоритма / В. С. Сидорова // Интерэкспо ГЕО-Сибирь. – 2011. – Т. 4. – С. 116-122.

74. Ester, M. A density-based algorithm for discovering clusters in large spatial database / M. Ester, H.-P. Kriegel, J. Sander, X. Xu // Proceedings of the 1996 International Conference on Knowledge Discovery and Data Mining (KDD'96). – 1996. – P. 226-231.
75. Sander, J. Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications / J. Sander, M. Ester, H.-P. Kriegel, X. Xu // Data Mining and Knowledge Discovery. – 1998. – Vol. 2. – No. 2. – P. 169-194.
76. Xu, X. A fast parallel clustering algorithm for large spatial databases / X. Xu, M. Ester, H.-P. Kriegel // Data Mining and Knowledge Discovery. – 1999. – Vol. 3. – No. 3. – P. 263-290.
77. Ester, M. Incremental clustering for mining in a data warehousing environment / M. Ester, H. Kriegel, J. Sander et al. // Proceedings of the 24th International Conference on Very Large Data Bases (VLDB'98). – N.Y.: Morgan Kaufmann, 1998. – P. 323-333.
78. Kriegel, H.-P. Incremental OPTICS: Efficient computation of updates in a hierarchical cluster ordering / H.-P. Kriegel, P. Kröger, I. Gotlibovich // Proceedings of the 5th International Conference on Data Warehousing and Knowledge Discovery (DaWaK'03). – Prague, Czech Republic, 2003. – P. 224-233.
79. Brecheisen, S. Parallel density-based clustering of complex objects / S. Brecheisen, H.-P. Kriegel, M. Pfeifle // Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'06). – Singapore, 2006. – P. 179-188.
80. Achtert, E. DeLiClu: Boosting robustness, completeness, usability, and efficiency of hierarchical clustering by a closest pair ranking / E. Achtert, C. Bohm, P. Kroger // Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'06). – Singapore, 2006. – P. 119-128.
81. Kroger, P. Density-connected subspace clustering for high-dimensional data / P. Kroger, H.-P. Kriegel, K. Kailing // Proceedings of the 4th SIAM International Conference on Data Mining. – Lake Buena Vista, FL, 2004. – P. 246-257.

82. Dash, M. '1+1>2': merging distance and density based clustering / M. Dash, H. Liu, X. Xu // Proceedings of the Seventh International Conference on Database Systems for Advanced Applications. – Hong-Kong: IEEE Computer Society, 2001. – P. 32-39.
83. Yanchang, Z. GDILC: A Grid-based density-isoline clustering algorithm / Z. Yanchang, S. Junde // Proceedings of the International Conference on Info-tech and Info-net (ICII 2001). – Beijing, China, 2001. – Vol. 3. – P. 140-145.
84. Zhao, Y. AGRID: An efficient algorithm for clustering large high-dimensional datasets / Y. Zhao, J. Song // Proceedings of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining. – Seoul, Korea, 2003. – P. 271-282.
85. Zhao, Y. Enhancing grid-density based clustering for high dimensional data / Y. Zhao, J. Cao, C. Zhang, S. Zhang // The Journal of Systems and Software. – 2011. – Vol. 84. – No. 9. – P. 1524-1539.
86. Ding, J. densityCut: an efficient and versatile topological approach for automatic clustering of biological data / J. Ding, S. Shah, A. Condon // Bioinformatics. – 2016. – Vol. 32. – No. 17. – P. 2567-2576.
87. Parzen, E. On the estimation of a probability density function and the mode / E. Parzen // The Annals of Mathematical Statistics. – 1962. – Vol. 33. – P. 1065-1076.
88. Rosenblatt, M. Remarks on some nonparametric estimates of a density function / M. Rosenblatt // The Annals of Mathematical Statistics. – 1956. – Vol. 27. – P. 832-837.
89. Фукунага, К. Введение в статистическую теорию распознавания образов / К. Фукунага. – М.: Наука, 1979. – 368 с.
90. Hinneburg, A. A general approach to clustering in large databases with noise / A. Hinneburg, D. Keim // Knowledge and Information Systems. – 2003. – Vol. 5. – No. 4. – P. 387-415.
91. Fukunaga, K. The estimation of the gradient of a density function, with applications in patter recognition / K. Fukunaga, L. D. Hosteeler // IEEE Transactions on Informational Theory. – 1975. – Vol. 21. – P. 32-40.

92. Rodriguez, A. Clustering by fast search and find of density peaks / A. Rodriguez, A. Laio // Science. – 2014. – Vol. 344. – No. 6191. – P. 1492-1496.
93. Zhou, Z. Robust clustering by identifying the veins of clusters based on kernel density estimation / Z. Zhou, G. Si, Y. Zhang, K. Zheng // Knowledge-Based Systems. – 2018. – Vol. 159. – P. 309-320.
94. Xu, X. A distribution-based clustering algorithm for mining in large spatial databases / X. Xu, M. Ester, H.-P. Kriegel, J. Sander // Proceedings of the IEEE International Conference on Data Engineering. – 1998. – P. 324-331.
95. Berkhin, P. Survey of clustering data mining techniques: Technical report [Electronic resource] / Berkhin P. – Accrue Software, 2002. – 56 p. – URL: https://www.cc.gatech.edu/~isbell/reading/papers/berkhin02survey.pdf.
96. Wang, W. STING: A statistical information grid approach to spatial data mining / W. Wang, J. Yang, M. Muntz // Proceedings of the 1997 International Conference on Very Large Data Bases (VLDB'97). – 1997. – P. 186-195.
97. Sheikholeslami, G. WaveCluster: A multi-resolution clustering approach for very large spatial databases / G. Sheikholeslami, S. Chatterjee, A. Zhang // Proceedings of the 24th Conference on VLDB. – NY, 1998. – P. 428-439.
98. Barbara, D. Using the fractal dimension to cluster datasets / D. Barbara, P. Chen // Proceedings of the 6th ACM SIGKDD. – Boston, MA, 2000. – P. 260-264.
99. Chang, C.-I. An axis-shifted grid-clustering algorithm / C.-I. Chang, N. P. Lin, N.-Y. Jan // Tamkang Journal of Science and Engineering. – 2009. – Vol. 12. – No. 2. – P. 183-192.
100. Shi, Y. A shrinking-based clustering approach for multidimensional data / Y. Shi, Y. Song, A. Zhang // IEEE Transactions on Knowledge and Data Engineering. – 2005. – Vol. 17. – No. 10. – P. 1389-1403.
101. Ma, E. W. M. A new shifting grid clustering algorithm / E. W. M. Ma, T. W. S. Chow // Pattern Recognition. – 2004. – Vol. 37. – P. 503-514.
102. Agrawal, R. Automatic subspace clustering of high dimensional data for data mining applications / R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan // SIGMOD Record ACM Special Interest Group on Management of Data. – 1998. – P. 94-105.

103. Cheng, C.-H. Entropy-based subspace clustering for mining numerical data / C.-H. Cheng, A. W. Fu, Y. Zhang // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – ACM Press, 1999. – P. 84-93.
104. Schikuta, E. Grid-clustering: A hierarchical clustering method for very large data sets / E. Schikuta // Proceedings of the 13th International Conference on Pattern Recognition. – 1993. – Vol. 2. – P. 101-105.
105. Pilevar, A. GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases / A. Pilevar, M. Sukumar // Pattern Recognition Letters. – 2005. – Vol. 26. – No. 7. – P. 999-1010.
106. Qiu, B.-Z. Grid-based clustering algorithm based on intersecting partition and density estimation / B.-Z. Qiu, X.-L. Li, J.-Y. Shen // Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. – Springer, Berlin, Heidelberg, 2007. – P. 368-377.
107. Nagesh, H. Adaptive grids for clustering massive data sets [Electronic resource] / H. Nagesh, S. Goil, A. Choudhary // Proceedings of the 1st SIAM International Conference on Data Mining. – Chicago, IL, 2001. – 17 p. – URL: https://epubs.siam.org/doi/pdf/10.1137/1.9781611972719.7.
108. Goil, S. Mafia: Efficient and scalable subspace clustering for very large data sets: Technical report CPDC-TR-9906-010 [Electronic resource] / S. Goil, H. Nagesh, A. Choudhary. – Center for Parallel and Distributed Computing, Department of Electrical & Computer Engineering, North-western University, 1999. – 20 p. – URL: http://www.quretec.com/u/vilo/edu/2003-04/DM_seminar_2003_II/ver1/ /P12/articles/goil99mafia.pdf.
109. Nagesh, H. A scalable parallel subspace clustering algorithm for massive data sets / H. Nagesh, S. Goil, A. Choudhary // International Conference on Parallel Processing. – 2000. – P. 477-484.
110. Чулюков, В. А. Системы искусственного интеллекта. Практический курс: Учебное пособие / В. А. Чулюков, И. Ф. Астахова, А. С. Потапов и др. – Москва: Бином, 2008. – 292 с.

111. Aljalbout, E. Clustering with deep learning: Taxonomy and new methods [Electronic resource] / E. Aljalbout, V. Golkov, Y. Siddiqui, M. Strobel, D. Cremers // arXiv preprint. – 2018. – URL: https://arxiv.org/pdf/1801.07648.pdf.
112. Lin, W. C. Constraint satisfaction neural networks for image segmentation / W. C. Lin, E. C. K. Tsao, C. T. Chen // Pattern Recognition. – 1992. – Vol. 25. – P. 679-693.
113. Cheng, K.-S. The application of competitive Hopfield neural network to medical image segmentation / K.-S. Cheng, J.-S. Lin, C.-W. Mao // IEEE Transactions on Medical Imaging. – 1996. – Vol. 15. – No. 4. – P. 560-567.
114. Pestunov, I. A. Algoriythms for processing polizonal video information for detection and classification of forests infested with insects / I. A. Pestunov // Pattern Recognition and Image Analysis. – 2001. – Vol. 11. – No. 2. – P. 368-371.
115. Freedman, D. Fast mean shift by compact density representation / D. Freedman, P. Kisilev // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2009. – P. 1818-1825.
116. Пестунов, И. А. Ансамблевый алгоритм кластеризации больших массивов данных / И. А. Пестунов, В. Б. Бериков, Е. А. Куликова, С. А. Рылов // Автометрия. – 2011. – Т. 47. – № 3. – С. 49-58.
117. Пестунов, И. А. Сеточный алгоритм кластеризации с использованием ансамблевого подхода к принятию решений / И. А. Пестунов, Е. А. Куликова, В. Б. Бериков, И. Д. Махатков // Вычислительные технологии. – Отдельный выпуск «Кузбасс 2». – 2009. – С. 52-64.
118. Епанечников, В. А. Непараметрическая оценка многомерной плотности вероятности / В. А. Епанечников // Теория вероятностей и ее применение. – 1969. – Т. 14. – № 1. – С. 156-160.
119. Comaniciu, D. Mean shift: A robust approach toward feature space analysis / D. Comaniciu, P. Meer // IEEE Transactions on Pattern Analysis Machine Intelligence. – 2002. – Vol. 24. – No. 5. – P. 603-619.
120. Comaniciu, D. Distribution free decomposition of multivariate data / D. Comaniciu, P. Meer // Pattern Analysis and Applications. – 1999. – Vol. 2. – P. 22-30.

121. Cheng, Y. Mean shift, mode seeking, and clustering / Y. Cheng // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1995. – Vol. 17. – P. 790-799.
122. Li, X. A note on the convergence of the mean shift / X. Li, Z. Hu, F. Wu // Pattern Recognition. – 2007. – Vol. 40. – P. 1756-1762.
123. Comaniciu, D. The variable bandwith mean shift and data-driven scale selection / D. Comaniciu, V. Ramesh, P. Meer // Proceedings of the Eighth IEEE International Conference on Computer Vision. – Vancouver, 2001. – Vol. 1. – P. 438-445.
124. Terrell, G. R. Variable kernel density estimation / G. R. Terrell, D. W. Scott // The Annals of Statistics. – 1992. – Vol. 20. – No. 3. – P. 1236-1265.
125. Rudzkis, R. On local bandwith selection for density estimation / R. Rudzkis, M. Kavaliauskas // Informatica. – 1998. – Vol. 9. – No. 4. – P. 479-490.
126. Comaniciu, D. An algorithm for data-driven bandwidth selection / D. Comaniciu // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2003. – Vol. 25. – No. 2. – P. 281-288.
127. Bugeau, A. Bandwidth selection for kernel estimation in mixed multi-dimension spaces [Electronic resource] / A. Bugeau, P. Pérez // E-print arXiv:0709.1920. – 2007. – 26 p. – URL: http://arxiv.org/pdf/0709.1920v2.pdf.
128. Wang, Y. Unsupervised color-texture segmentation based on soft criterion with adaptive mean-shift clustering / Y. Wang, J. Yang, N. Peng // Pattern Recognition Letters. – 2006. – Vol. 27. – P. 386-392.
129. Silverman, B. W. Density estimation for statistical and data analysis / B. W. Silverman. – London, N.Y.: Chapman and Hall, 1986. – 176 p.
130. Raykar, V. C. Very fast optimal bandwidth selection for univariate kernel density estimation: Technical report CS-TR-4774 [Electronic resource] / V. C. Raykar, R. Duraiswami. – University of Maryland, CollegePark, 2005. – 36 p. – URL: https://pdfs.semanticscholar.org/49c7/51944b30e8aefe7be5577c287e0fcad9d7a9.pdf.
131. Wang, B. Bandwidth selection for weighted kernel density estimation [Electronic resource] / B. Wang, X. Wang // Electronic Journal of Statistics. – 2007. – 22 p. – URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.244.7828&rep=rep1&type=pdf.

132. Pauwels, E. Finding salient regions in images / E. Pauwels, G. Frederix // Computer Vision and Image Understanding. – 1999. – Vol. 75. – P. 73-85.
133. Cornuéjols, A. Collaborative clustering: Why, when, what and how / A. Cornuéjols, C. Wemmert, P. Gançarski, Y. Bennani // Information Fusion. – 2018. – Vol. 39. – P. 81-95.
134. Liu, H. Infinite ensemble clustering / H. Liu, M. Shao, S. Li, Y. Fu // Data Mining and Knowledge Discovery. – 2018. – Vol. 32. – No. 2. – P. 385-416.
135. Fern, X. Z. Clustering ensembles for high dimensional data clustering / X. Z. Fern, C. E. Brodley // Proceedings of the International Conference on Machine Learning. – 2003. – P. 186-193.
136. Fern, X. Z. Clustering ensembles for high dimensional data clustering: An empirical study: Technical report CS06-30-02. [Electronic resource] / X. Z. Fern, C. E. Brodley. – 2004. – 26 p. – URL: http://web.engr.oregonstate.edu/~xfern/clustensem.pdf.
137. Fred, A. Combining multiple clusterings using evidence accumulation / A. Fred, A. K. Jain // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2005. – Vol. 27. – P. 835-850.
138. Strehl, A. Clustering ensembles – a knowledge reuse framework for combining multiple partitions / A. Strehl, J. Ghosh // The Journal of Machine Learning Research. – 2002. – Vol. 3. – P. 583-617.
139. Бериков, В. Б. Построение ансамбля деревьев решений в кластерном анализе / В. Б. Бериков, И. А. Пестунов // Вычислительные технологии. – 2016. – Т. 21. – № 1. – С. 15-24.
140. Журавлёв, Ю. И. Распознавание. Математические методы. Программная система. Практические применения / Ю. И. Журавлёв, В. В. Рязанов, О. В. Сенько. – М.: ФАЗИС, 2006. – 176 с.
141. Schapire, R. E. The boosting approach to machine learning: An overview [Electronic resource] / R. E. Schapire // MSRI Workshop on Nonlinear Estimation and Classification. – Springer, 2003. – 23 p. – URL: https://www.cs.princeton.edu/courses/archive/spring07/cos424/papers/boosting-survey.pdf.

142. Breiman, L. Bagging predictors / L. Breiman // Machine Learning. – 1996. – Vol. 24. – P. 123-140.

143. Hong, Y. To combine steady-state genetic algorithm and ensemble learning for data clustering / Y. Hong, S. Kwong // Pattern Recognition Letters. – 2008. – Vol. 29(9). – P. 1416-1423.

144. Asuncion, A. UCI machine learning repository [Electronic resource] / A. Asuncion, D. J. Newman. – Irvine, CA: University of California, School of Information and Computer Science, 2007. – URL: http://www.ics.uci.edu/~mlearn/X9MLRepository.html.

145. Кендал, М. Многомерный статистический анализ и временные ряды / М. Кендал, А. Стьюарт. – Москва: Наука. 1976. – С. 441-443.

146. Куликова, Е. А. Непараметрический алгоритм кластеризации для обработки больших массивов данных / Е. А. Куликова, И. А. Пестунов, Ю. Н. Синявский // Сборник докладов XIV всероссийской конференции «Математические методы распознавания образов». – Москва, 21-26 сентября 2009 г. – Москва: МАКС Пресс, 2009. – С. 149-152.

147. Пестунов, И. А. Классификация больших массивов данных в условиях малой априорной информации / И. А. Пестунов, Д. И. Добротворский, Ю. Н. Синявский // Вычислительные технологии. – 2007. – Т. 12. – Спецвыпуск 4. – С. 50-58.

148. Куликова, Е. А. Классификация с полуобучением в задачах обработки многоспектральных изображений / Е. А. Куликова, И. А. Пестунов // Вычислительные технологии. – 2008. – Т. 13. – Вестник КазНУ им. аль-Фараби. Серия: Математика, механика, информатика. – 2008. – № 3(58). – Совместный выпуск. – Ч. II. – С. 284-291.

149. Шокин, Ю. И. Корпоративная информационная система СО РАН сбора, хранения и обработки спутниковых данных / Ю. И. Шокин, И. А. Пестунов, В. В. Смирнов и др. // Горный информационно-аналитический бюллетень. – 2009. – Отдельный выпуск «Кузбасс 2». – С. 3-10.

150. Пестунов, И. А. Каталог пространственных данных для решения задач регионального мониторинга / И. А. Пестунов, В. В. Смирнов, О. Л. Жижимов, Ю. Н. Синявский, А. П. Скачкова, И. С. Дубров // Вычислительные технологии. – 2008. – Т. 13. – Вестник КазНУ им. аль-Фараби. Серия: Математика, механика, информатика. – 2008. – № 4 (59). – Совместный выпуск. – Ч. III. – С. 71-76.

151. Гопп, Н. В. Разделение формаций растительности с близкими спектрально-яркостными характеристиками по данным съемки со спутника Landsat 7 ETM+ / Н. В. Гопп, Е. А. Куликова, И. А. Пестунов, В. В. Смирнов, Ю. Н. Синявский // Вычислительные Технологии. – 2007. – Т. 12. – Спецвыпуск 2. – С. 194-201.

152. Chavez, P. S. Statistical method for selecting Landsat MSS ratios / P. S. Chavez, G. L. Berlin, L. B. Sowers // Journal of Applied Photographic Engineering. – 1984. – Vol. 8. – P. 23-30.

153. Пестунов, И. А. Обнаружение и картирование повреждений кедровых древостоев по изображениям со спутника Pleiades / И. А. Пестунов, П. В. Мельников, О. А. Дубровская, Ю. Н. Синявский, В. И. Харук // Интерэкспо Гео-Сибирь. – 2014. – Т. 3. – № 2. – С. 400-408.

154. Дейвис, Ш. М. Дистанционное зондирование: количественный подход. Перевод с английского / Ш. М. Дейвис, Д. А. Ландгребе, Т. Л. Филипс и др. – Москва: Недра, 1983. – 415 с.

ПУБЛИКАЦИИ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИОННОЙ РАБОТЫ

В рецензируемых журналах, рекомендованных ВАК:

1. Пестунов, И. А. Непараметрический алгоритм кластеризации данных дистанционного зондирования на основе grid-подхода / И. А. Пестунов, Ю. Н. Синявский // Автометрия. – 2006. – Т. 42. – № 2. – С. 90-99.
2. Пестунов, И. А. Сегментация многоспектральных изображений на основе ансамбля непараметрических алгоритмов кластеризации / И. А. Пестунов, В. Б. Бериков, Ю. Н. Синявский // Вестник Сибирского государственного аэрокосмического университета. – 2010. – Выпуск 5 (31). – С. 56-64.
3. Пестунов, И. А. Алгоритмы кластеризации в задачах сегментации спутниковых изображений / И. А. Пестунов, Ю. Н. Синявский // Вестник Кемеровского государственного университета. – 2012. – Т. 52. – № 4/2. – С. 110-125.
4. Шокин, Ю. И. Распределенная информационная система сбора, хранения и обработки спутниковых данных для мониторинга территорий Сибири и Дальнего Востока / Ю. И. Шокин, И. А. Пестунов, В. В. Смирнов, Ю. Н. Синявский и др. // Журнал Сибирского федерального университета. Техника и технологии. – 2008. – Т. 1. – Выпуск 4. – С. 291-314.
5. Синявский, Ю. Н. Методы и технология сегментации мультиспектральных изображений высокого разрешения для исследования природных и антропогенных объектов / Ю. Н. Синявский, И. А. Пестунов, О. А. Дубровская, С. А. Рылов, П. В. Мельников, Н. Б. Ермаков, М. А. Полякова // Вычислительные технологии. – 2016. – Т. 21. – № 1. – С. 127-140.
6. Смирнов, В. В. Корпоративные картографические сервисы Сибирского отделения РАН / В. В. Смирнов, И. А. Пестунов, Д. И. Добротворский, Ю. Н. Синявский // Горный информационно-аналитический бюллетень. – 2009. – № S18. – С. 130-135.
7. Шокин, Ю. И. Корпоративная информационная система СО РАН для сбора, хранения и обработки спутниковых и наземных данных / Ю. И. Шокин, И. А. Пестунов, В. В. Смирнов, Ю. Н. Синявский, Д. И. Добротворский,

А. П. Скачкова // Горный информационно-аналитический бюллетень. – 2009. – № S17. – С. 9-15.

В изданиях, входящих в международные базы данных:

8. Sinyavskiy, Yu. N. Extension of training set using mean shift procedure for aerospace images classification / Yu. N. Sinyavskiy, P. V. Melnikov, I. A. Pestunov // E3S Web of Conferences. 2019. – Vol. 75. – No 14. – Article no. 01010. – 2018 Regional Problems of Earth Remote Sensing, RPERS 2018. – Krasnoyarsk, Russian Federation. 11 Sept. 2018 – 14 Sept. 2018. (WoS, Scopus)
9. Pestunov, I. A. Non-parametric grid-based clustering algorithm for remote sensing data / I. A. Pestunov, Yu. N. Sinyavsky // Proceedings of the Second IASTED International Multi-Conference on Automation, Control, and Information Technology. – Novosibirsk, Russia, 2005. – P. 5-9. (WoS, Scopus)
10. Pestunov, I. A. Computationally efficient methods of clustering ensemble construction for satellite image segmentation / I. A. Pestunov, S. A. Rylov, Yu. N. Sinyavskiy, V. B. Berikov // CEUR Workshop Proceedings – Proceedings of the International Conference on Information Technology and Nanotechnology. Session Image Processing, Geoinformation Technology and Information Security. – 2017. – P. 194-200. (Scopus)

В рецензируемых журналах:

11. Гопп, Н. В. Разделение формаций растительности с близкими спектрально-яркостными характеристиками по данным съемки со спутника Landsat 7 ETM+ / Н. В. Гопп, Е. А. Куликова, И. А. Пестунов, Ю. Н. Синявский, В. В. Смирнов // Вычислительные технологии. – 2007. – Т. 12. – Спецвыпуск 2. – С. 194-201.
12. Пестунов, И. А. Непараметрический алгоритм кластеризации многоспектральных аэрокосмических данных, основанный на процедуре «среднего сдвига» / И. А. Пестунов, Ю. Н. Синявский // Вычислительные технологии. – 2004. – Т. 9. – Спецвыпуск. – Труды Совещания российско-казахстанской рабочей группы по вычислительным и информационным технологиям. – С. 125-132.

13. Пестунов, И. А. Обнаружение и картирование повреждений кедровых древостоев по изображениям со спутника Pleiades / И. А. Пестунов, П. В. Мельников, О. А. Дубровская, Ю. Н. Синявский, В. И. Харук // Интерэкспо Гео-Сибирь. – 2014. – Т. 3. – № 2. – С. 400-408.
14. Синявский, Ю. Н. Экспериментальное сравнение непараметрических алгоритмов кластеризации для сегментации мультиспектральных изображений / Ю. Н. Синявский, С. А. Рылов // Интерэкспо Гео-Сибирь. – 2018. – Т. 1. – № 4. – С. 109-114.
15. Добротворский, Д. И. Веб-сервисы для непараметрической классификации спутниковых данных / Д. И. Добротворский, Е. А. Куликова, И. А. Пестунов, Ю. Н. Синявский // Гео-Сибирь. – 2010. – Т. 1. – № 2. – С. 171-175.
16. Синявский, Ю. Н. Совместная обработка разнородных данных при сегментации спутниковых изображений высокого разрешения / Ю. Н. Синявский, И. А. Пестунов, С. А. Рылов, П. В. Мельников // Интерэкспо Гео-Сибирь. – 2015. – Т. 4. – № 2. – С. 57-61.
17. Пестунов, И. А. Технология и программный инструментарий для сегментации спутниковых изображений высокого пространственного разрешения / И. А. Пестунов, С. А. Рылов, П. В. Мельников, Ю. Н. Синявский // Интерэкспо Гео-Сибирь. – 2013. – Т. 4. – № 1. – С. 202-208.
18. Шокин, Ю. И. Система сбора, хранения и обработки данных дистанционного зондирования для исследования территорий Западной и Восточной Сибири / Ю. И. Шокин, И. А. Пестунов, В. В. Смирнов, Ю. Н. Синявский, А. П. Скачкова, И. С. Дубров // Гео-Сибирь. – 2009. – Т. 4. – № 1. – С. 165-170.

В трудах международных и всероссийских конференций:

19. Синявский, Ю. Н. Методы и технология сегментации мультиспектральных изображений высокого разрешения для исследования природных и антропогенных объектов [Электронный ресурс] / Ю. Н. Синявский, И. А. Пестунов, С. А. Рылов, П. В. Мельников // Сборник трудов всероссийской конференции «Обработка пространственных данных в задачах мониторинга природных и антропогенных процессов» (24-28 августа 2015 г.,

с. Усть-Сема, Республика Алтай). – Новосибирск: ИВТ СО РАН, 2015. – С. 107-114. – Адрес доступа: http://conf.nsc.ru/files/conferences/SDM-2015/294652/SDM-2015%20Thesis.pdf.

20. Синявский, Ю. Н. Использование разнородных данных при сегментации спутниковых изображений высокого разрешения / Ю. Н. Синявский, И. А. Пестунов, О. А. Дубровская, П. В. Мельников, С. А. Рылов, Д. В. Лазарев // Материалы международной конференции «Вычислительные и информационные технологии в науке, технике и образовании (CITech-2015)». – Алмата, Казахстан, 2015. – С. 316-323.
21. Пестунов, И. А. Подход к построению ансамбля непараметрических алгоритмов кластеризации для сегментации спутниковых изображений / И. А. Пестунов, С. А. Рылов, Ю. Н. Синявский, В. Б. Бериков // Сборник трудов III международной конференции и молодежной школы «Информационные технологии и нанотехнологии (ИТНТ-2017)». – Самара: Самарский национальный Исследовательский университет им. академика С. П. Королева, 2017. – С. 775-780.
22. Синявский, Ю. Н. Наращивание обучающей выборки с помощью процедуры «среднего сдвига» в задачах классификации спутниковых изображений / Ю. Н. Синявский, П. В. Мельников, И. А. Пестунов // Материалы международной научной конференции «Региональные проблемы дистанционного зондирования Земли». – Красноярск: Сибирский федеральный университет, Институт космических и информационных технологий, 2018. – С. 211-215.

В тезисах международных и всероссийских конференций:

23. Синявский, Ю. Н. Сервис-ориентированный подход к обработке спутниковых изображений / Ю. Н. Синявский, И. А. Пестунов // Тезисы XIV Всероссийской конференции молодых ученых по математическому моделированию и информационным технологиям. – Новосибирск: ИВТ СО РАН, 2013. – С. 43-44.
24. Пестунов, И. А. Ансамблевые алгоритмы сегментации мультиспектральных

спутниковых изображений высокого пространственного разрешения и их практическое применение [Электронный ресурс] / И. А. Пестунов, П. В. Мельников, С. А. Рылов, Ю. Н. Синявский, С. Т. Им, В. И. Харук // Сборник тезисов докладов XI Всероссийской открытой конференции «Современные проблемы дистанционного зондирования Земли из космоса». – Москва: ИКИ РАН, 2013. – Адрес доступа: http://smiswww.iki.rssi.ru/d33_conf/thesisshow.aspx?page=78&thesis=3909

25. Пестунов, И. А. Технология сегментации многоспектральных спутниковых изображений высокого пространственного разрешения / И. А. Пестунов, Ю. Н. Синявский, П. В. Мельников, С. А. Рылов // Тезисы Всероссийской конференции «Обработка пространственных данных и дистанционный мониторинг природной среды и масштабных антропогенных процессов» (DPRS-2013). – Барнаул: Пять плюс, 2013. – С. 55-56.
26. Пестунов, И. А. Применение ансамблей непараметрических алгоритмов кластеризации для обработки многоспектральных спутниковых изображений [Электронный ресурс] / И. А. Пестунов, Е. А. Куликова, Ю. Н. Синявский, В. В. Смирнов // Тезисы Восьмой открытой Всероссийской конференции «Современные проблемы дистанционного зондирования земли из космоса (Физические основы, методы и технологии мониторинга окружающей среды, потенциально опасных явлений и объектов)». – Москва, 15-19 ноября 2010 г. – Москва: ИКИ РАН, 2010. – С. 40-41. – Адрес доступа: http://d902.iki.rssi.ru/theses-cgi/thesis.pl?id=2180.
27. Пестунов, И. А. Построение ансамблей сеточно-плотностных алгоритмов для кластеризации больших массивов данных в условиях малой априорной информации / И. А. Пестунов, В. Б. Бериков, Ю. Н. Синявский, Е. А. Куликова // Тезисы II Международной конференции «Геоинформатика: Технологии, научные проекты». – Барнаул: ООО «А.Р.Т.», 2010. – С. 79.
28. Синявский, Ю. Н. MeanSC: Непараметрический алгоритм кластеризации данных дистанционного зондирования / Ю. Н. Синявский // Тезисы докладов VII Всероссийской конференции молодых ученых по математическому

моделированию и информационным технологиям. – Красноярск, 1-3 ноября 2006 г. – Красноярск. 2006. – С. 95.

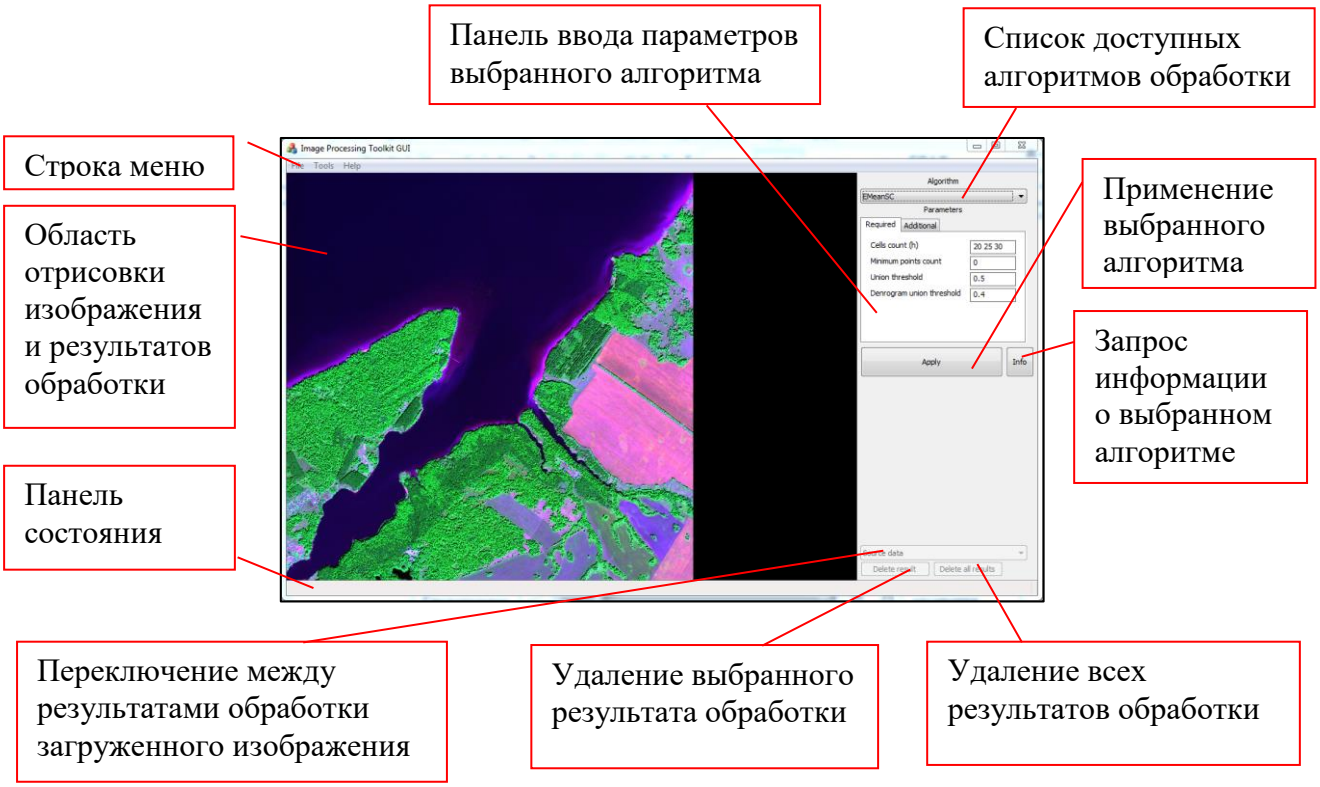
29. Синявский, Ю. Н. Быстрый непараметрический алгоритм классификации многоспектральных аэрокосмических данных / Ю. Н. Синявский // Материалы XLIII Международной научной студенческой конференции «Студент и научно-технический прогресс». – Новосибирск, 2005. – С. 134.

Зарегистрированные программы для ЭВМ:

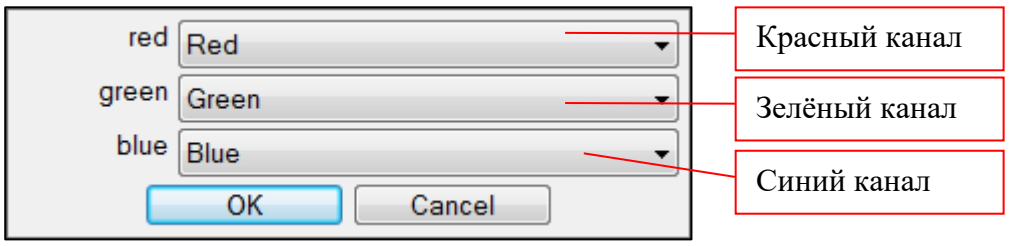
1. Синявский, Ю. Н. MeanSC / Ю. Н. Синявский, И. А. Пестунов // Свидетельство о государственной регистрации программы для ЭВМ № 2016617547 от 07.07.2016 г. (Федеральная служба по интеллектуальной собственности, патентам и товарным знакам).
2. Синявский, Ю. Н. Программа для сегментации изображений «IP_EMeanSC» / Ю. Н. Синявский, И. А. Пестунов // Свидетельство о государственной регистрации программы для ЭВМ № 2019664225 от 01.11.2019 г. (Федеральная служба по интеллектуальной собственности, патентам и товарным знакам).
3. Синявский, Ю. Н. Программа для обработки спутниковых изображений «Image Processing Toolkit» / Ю. Н. Синявский // Свидетельство о государственной регистрации программы для ЭВМ № 2019615674 от 06.05.2019 г. (Федеральная служба по интеллектуальной собственности, патентам и товарным знакам).
4. Синявский, Ю. Н. Программа «Image Processing Toolkit GUI» / Ю. Н. Синявский // Свидетельство о государственной регистрации программы для ЭВМ № 2019615205 от 22.04.2019 г. (Федеральная служба по интеллектуальной собственности, патентам и товарным знакам).
5. Синявский, Ю. Н. Программа для наращивания обучающей выборки в задачах классификации спутниковых изображений «IP_SPMSROIExtension» / Ю. Н. Синявский, И. А. Пестунов // Свидетельство о государственной регистрации программы для ЭВМ № 2019664226 от 01.11.2019 г. (Федеральная служба по интеллектуальной собственности, патентам и товарным знакам).

ПРИЛОЖЕНИЕ 1. ГРАФИЧЕСКИЕ ПОЛЬЗОВАТЕЛЬСКИЕ ИНТЕРФЕЙСЫ ПАКЕТА «IMAGE PROCESSING TOOLKIT»

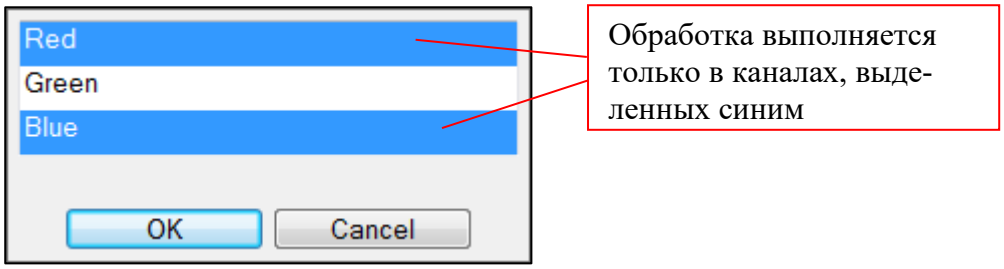
Основное окно программы



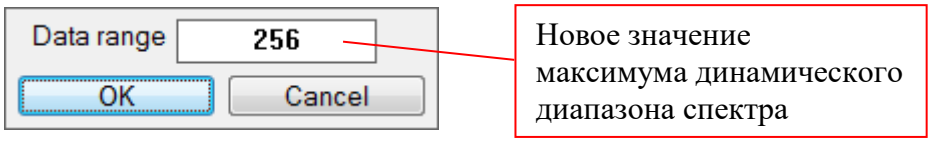
Диалог выбора каналов для отображения



Диалог выбора каналов для обработки



Диалог линейного растяжения динамических диапазонов спектра



ПРИЛОЖЕНИЕ 2. СВИДЕТЕЛЬСТВА О ГОСУДАРСТВЕННОЙ РЕГИСТРАЦИИ ПРОГРАММ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2016617547

MeanSC

Правообладатель: ***Федеральное государственное бюджетное учреждение науки Институт вычислительных технологий Сибирского отделения Российской академии наук (RU)***

Авторы: ***Синявский Юрий Николаевич (RU),***
Пестунов Игорь Алексеевич (RU)

Заявка № **2016612347**
Дата поступления **18 марта 2016 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***07 июля 2016 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Г.П. Ивлиев

Свидетельство о государственной регистрации программы MeanSC

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019664225

Программа для сегментации изображений «IP_EMeanSC»

Правообладатель: ***Федеральное государственное бюджетное учреждение науки Институт вычислительных технологий Сибирского отделения Российской академии наук (ИВТ СО РАН) (RU)***

Авторы: ***Синявский Юрий Николаевич (RU),***
Пестунов Игорь Алексеевич (RU)

Заявка № **2019663104**
Дата поступления **22 октября 2019 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***01 ноября 2019 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Г.П. Ивлиев

Свидетельство о государственной регистрации программы EMeanSC

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019615674

Программа для обработки спутниковых изображений «Image Processing Toolkit»

Правообладатель: ***Федеральное государственное бюджетное учреждение науки Институт вычислительных технологий Сибирского отделения Российской академии наук (ИВТ СО РАН) (RU)***

Автор: ***Синявский Юрий Николаевич (RU)***

Заявка № **2019613857**
Дата поступления **10 апреля 2019 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***06 мая 2019 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

Свидетельство о государственной регистрации программы Image Processing Toolkit

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019615205

Программа «Image Processing Toolkit GUI»

Правообладатель: ***Федеральное государственное бюджетное учреждение науки Институт вычислительных технологий Сибирского отделения Российской академии наук (ИВТ СО РАН) (RU)***

Автор: ***Синявский Юрий Николаевич (RU)***

Заявка № **2019614111**

Дата поступления **15 апреля 2019 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***22 апреля 2019 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

Свидетельство о государственной регистрации программы Image Processing Toolkit GUI

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019664226

Программа для наращивания обучающей выборки в задачах классификации спутниковых изображений «IP_SPMSROIExtension»

Правообладатель: ***Федеральное государственное бюджетное учреждение науки Институт вычислительных технологий Сибирского отделения Российской академии наук (ИВТ СО РАН) (RU)***

Авторы: ***Синявский Юрий Николаевич (RU),***
Пестунов Игорь Алексеевич (RU)

Заявка № **2019663093**
Дата поступления **22 октября 2019 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***01 ноября 2019 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

Свидетельство о государственной регистрации программы SPMSROIExtension

ПРИЛОЖЕНИЕ 3. АКТ ИСПОЛЬЗОВАНИЯ РЕЗУЛЬТАТОВ ДИССЕРТАЦИОННОЙ РАБОТЫ В ИПА СО РАН

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ

ИНСТИТУТ ПОЧВОВЕДЕНИЯ И АГРОХИМИИ
СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК
(ИПА СО РАН)

630090, Новосибирск 90, просп. Академика Лаврентьева, 8/2
Для телеграмм: Новосибирск 90, Почва
Тел/факс: (383) 36-39-025
E-mail: soil@issa-siberia.ru
ИНН 5406015286

_____ № 15343-

На № _____ от _____

Директор ИПА СО РАН, д.б.н.

В.А. Андроханов

АКТ
об использовании результатов исследований

Настоящим Актом подтверждается, что результаты исследований Синявского Юрия Николаевича, изложенные в диссертации на соискание учёной степени кандидата технических наук «Непараметрические методы и программно-алгоритмический инструментарий для сегментации мультиспектральных спутниковых изображений», были использованы сотрудниками Института почвоведения и агрохимии СО РАН при построении детальных картографических моделей. Методики, разработанные на основе алгоритма MeanSC, позволили разделить формации лесной растительности на территории Болотнинского района Новосибирской области с близкими спектрально-яркостными характеристиками по данным со спутника Landsat-7, а также выполнить картографическое моделирование структурной организации почвенного покрова тундрово-степных комплексов высокогорий Алтае-Саянского региона.

Старший научный сотрудник ИПА СО РАН, к.б.н. Кудряшова С.Я.

(подпись)