Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Комсомольский-на-Амуре государственный университет»

На правах рукописи

Намоконов Александр Николаевич

Модели формирования наборов данных для задач дистанционного зондирования

Направление подготовки 09.04.03
«Прикладная информатика»

АВТОРЕФЕРАТ
МАГИСТЕРСКОЙ ДИССЕРТАЦИИ

2022

Работа выполнена в ФГБОУ ВО
«Комсомольский-на-Амуре государственный университет»

Научный руководитель:	кандидат физико-математических наук, доцент кафедры «Прикладная математика» Григорьева Анна Леонидовна
Рецензент:	кандидат технических наук, доцент кафедры информационной безопасности, ФГБОУ ВО «АМГПГУ» Анисимов Антон Николаевич

Защита состоится 23 июня 2021 г. в 9.50 часов на заседании государственной экзаменационной комиссии по направлению подготовки 09.04.03 – «Прикладная информатика» в ФГБОУ ВО «КнАГУ» по адресу: 681000, г. Комсомольск-на-Амуре, пр. Ленина, 27, ауд. 204/5.

Автореферат разослан 21 июня 2022 г.

Секретарь ГЭК

ОБЩАЯ ХАРАКТЕРИСТИКА ДИССЕРТАЦИОННОЙ РАБОТЫ

Актуальность темы.

Машинное обучение быстро набирает обороты как в исследованиях, так и в метод прогнозировании сложные физических явлений. Науки о Земле особенно выигрывают от этих достижений в таких приложениях, как прогнозирование осадков, обнаружение лесных пожаров и сельскохозяйственное планирование.

Области использования пространственных изображений стали разнообразными. Источников с количеством изображений увеличилось, и доступ к ним значительно облегчился.

Материалы космической съёмки в нескольких зонах энергетического спектра, преимущественно в видимом (0,4–0,7 мкм) и ближнем инфракрасном (0,7–1,3 мкм) диапазонах, представляют большой интерес для решения многих задач. Спектральные отражательные свойства растительности и почвеннорастительных комплексов зависят от состава, структуры, фазы вегетации, климатических и многих других факторов. Поэтому в настоящее время всё больше и больше возрастает интерес к обработке данных, полученных со спутников дистанционного зондирования Земли (ДЗЗ). Эти данные могут использоваться для решения самых различных задач, таких как мониторинг состояния почвы, водоёмов и растительности; выявление очагов лесных пожаров; мониторинг и оценка эффективности лесовосстановительной деятельности; контроль природопользования (вырубки леса, строительства карьеров, незаконных свалок, оценка рациональности при добыче природных ресурсов и пр.); Применение различных методов предварительной обработки снимков позволяет сократить время, затраченное на решение задачи, а в некоторых случаях полностью автоматизируют этот процесс.

Предметной областью исследования являются спутниковые снимки

Объектом исследования являются образ карты местности

Цель дипломной работы состоит в определение возможностей оценки состояния поверхностей на основе мультиспектральных изображений искусственных спутников земли.

Для достижения цели необходимо решить следующие задачи:

- определение основного набора данных, позволяющих строить карту, содержащую n характеристик в каждой её точке;
- анализ полученной карты и выделение критических областей, соответствующих состоянию оцениваемой поверхности;

Теоретико-методологической основой работы методологической основой исследования выступают:

- Классификация объектов
- статистический анализ данных;

Практическая значимость заключается в построении модели, позволяющей, без постоянного участия человека, оценить состояние и в самые короткие сроки построить карту местности, содержащую описание критических областей.

Научная новизна исследования:

- проводится комплексный анализ существующих индексов, характеризующих состояние поверхности содержащих объекты живой природы, строится карта местности для определения типа исследуемой поверхности

Апробация работы. Основные положения дипломной работы докладывались на V Всероссийской национальной научной конференции молодых ученых «Молодежь и наука». По итогам конференции было опубликовано две статьи.

Структура и объем диссертации. Дипломная работа состоит из введения, пяти глав и заключения. Содержит 47 страниц, 17 рисунков. Список литературы состоит из 11 наименований.

Публикации

– молодежь и наука: актуальные проблемы фундаментальных и прикладных исследований материалы V Всероссийской. нац. науч. конф. студентов, аспирантов и молодых ученых, Комсомольск-на-Амуре, (Комсомольск-на-Амуре, апрель 2021 г.).

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Введение раскрывает актуальность темы, определяются цели и задачи исследования, объект, предмет, указываются научная новизна, практическая значимость, достоверность и обоснованность результатов исследования.

В первой главе описан основной подход к дистанционному зондированию Земли, приведены математические методы и модели для кластеризации изображений со спутника

Разработка методологии машинного обучения для дистанционного зондирования стала значительным шагом в развитии ДЗЗ сферы. Но с применением алгоритмов машинного обучения на данных ДЗЗ появились и новые нерешенные задачи.

Во время работы модели машинного обучения (ML), одновременно запускаются два процесса. Во-первых, тонны данных собираются со спутников ДЗЗ, которые обрабатываются, чтобы сделать их готовыми к применению. Эти данные называются готовыми к применению данными (ARD), помещаются в облако и организуются в различные наборы данных, называемые кубами данных. Во-вторых, данные обучения собираются для обучения моделей. Как только оба набора данных организованы, выбирается подходящая модель машинного обучения для классификации, сглаживания и обработки данных, чтобы получить ценную информацию.

Использование нескольких ML алгоритмов на больших объемах данных ДЗЗ обеспечивает надежные и более точные результаты, тем самым облегчая процесс доказательства или опровержения заданной гипотезы. Преимущества использования машинного обучения многочисленны, но общая доступность спутниковых данных ДЗЗ затрудняет эффективное использование моделей и алгоритмов ML. В настоящее время мы получаем тонну наборов данных с таких спутников как Sentinel 2, Sentinel 3, Landsat 8 и SkySat, которые предо-

ставляют более 2 петабайт (PB) данных каждый день. Таким образом, хотя многие ML модели эффективно работают на тестовых моделях, они не отражают реальную действительность.

Одной из наиболее важных проблем, с которыми приходится сталкиваться при надлежащем развертывании ML моделей, является огромный объем собранных данных. Однако здесь имеет решающее значение маркировка данных по количеству категорий. Это определяет поведение классификатора, моделирующего данные.

ML необходимо маркировать данные, чтобы лучше их понимать, но разнородность данных ограничивает применение алгоритмов ML. Существующая в настоящее время классификация часто оказывается недостаточной для обозначения данных.

Чтобы подвести итог, нужно измерить то, что существует в определенном месте в определенный момент времени, и определить события, которые произошли в этом конкретном месте с течением времени. Следовательно, для работы с пространством и временем нам нужны пространственно-временные модели.

Моделирование событий и времени имеет ключевое значение для анализа больших данных ДЗЗ, но машинному обучению трудно справиться с этими изменениями. Решение вышеприведенной проблемы заключается в использовании геопространственной семантики для анализа данных ДЗЗ.

Другие технические проблемы, с которыми сталкиваются аналитики данных и процессоры при подаче изображений в модели ML, такие как:

Разрешение – Различные спутники обеспечивают различное разрешение изображений в диапазоне от 500 м MODIS, до 0,3 м. Кроме того, различные наборы данных имеют различные форматы, такие как JPEG2000 и GeoTIFF. Таким образом, процессор должен научиться работать с различными разрешениями и форматами. Эта проблема может быть частично решена сторонним программным обеспечением, таким как Sentinel Hub, которое гармонизирует данные наблюдения Земли в одном единственном формате.

Проблема облачности – Несмотря на то, что облака имеют 5-дневный цикл повторного посещения, спутниковые снимки часто частично или полностью покрыты облаками. Облака затрудняют любому алгоритму и процессору получение полезной информации из спутниковых изображений. Поэтому процессор должен маскировать эти облака так, чтобы эти белые пятна или тени не искажали сигналы.

Сегментация является одним из важнейших этапов анализа цифровых изображений. Она заключается в разбиении изображения на непересекающиеся области на основе похожести их спектральных или пространственных характеристик (текстура, размер, форма и т.п.). Методы сегментации нашли широкое применение во многих прикладных областях, в том числе в дистанционном зондировании Земли, интерес к которому в последние годы непрерывно возрастает. Наиболее распространенный подход к сегментации спутниковых изображений основан на использовании алгоритмов кластеризации данных.

Кластеры определяются как области в пространстве признаков с высокой плотностью распределения данных, разделенные областями с низкой плотностью. Для непараметрического оценивания плотности часто используются гистограммная оценка и оценка Розенблатта–Парзена.

Известно, что устойчивость решений в задачах кластеризации может быть повышена благодаря использованию ансамблевого подхода, который является одним из наиболее перспективных направлений в кластерном анализе.

Ансамблевый подход заключается в формировании согласованного результата кластеризации основе нескольких вариантов разбиения данных. Комбинирование различных кластерных решений позволяет повысить качество результатов и их устойчивость к изменению параметров.

При построении ансамблевого алгоритма кластеризации требуется решить два ключевых вопроса – как получить разнообразные разбиения и каким образом их согласовать. Существует четыре основных способа получения начальных разбиений: использование различных алгоритмов кластеризации; использование некоторого алгоритма кластеризации с разными параметрами; выбор различных подмножеств признаков и использование различных подмножеств данных.

Одним из наиболее эффективных методов построения ансамблевого решения является использование согласованной матрицы попарного сходства/различия (co-association matrix). Элементы этой матрицы характеризуют попарную схожесть объектов как количество разбиений, в которых эти объекты относятся к одному кластеру. Для получения итогового решения матрица различий используется как матрица расстояний между объектами. К ней применяется один из стандартных иерархических алгоритмов кластеризации. Данный метод не требует совпадения количества кластеров во всех разбиениях. Это условие необходимо для непараметрической кластеризации, когда число получаемых кластеров заранее не определено.

В кластерном анализе требуется получить разбиение

$P = \{C1,\dots,CK\}$

множества некоторых элементов (объектов) $A = \{a_1,\dots,a_N\}$

на определенное число K групп (класте-ров) в соответствии с заданным критерием качества. Под критерием качества понимается некоторый функционал, зависящий от разброса внутри группы и расстояний между группами. Как правило, каждый объект описывается с помощью набора вещественных переменных $X_1,\dots,X_n$.

Через $x = x(a) = (x1,\dots,xn)$ обозначим вектор переменных

для объекта a, где $xj = \boldsymbol{X}j(a), j = 1,\dots,n,$ а через T_{Nxn} — матрицу (таблицу данных)

$\left(x(a_1),\dots,x(a_N)\right)^T$

Число кластеров может быть задано или не задано;

В анализе изображений под элементом понимается пиксель, а переменные описывают различные свойства пикселей (спектральную яркость в задан-

ном диапазоне, текстурные характеристики и т. д.). Например, RGB-изображение может быть представлено в виде таблицы данных с помощью трех переменных X_1, X_2, X_3, характеризующих интенсивность соответственно красной, зеленой и синей составляющих цвета каждого пикселя. Для гиперспектрального изображения каждый пиксель может быть охарактеризован упорядоченной последовательностью $X_1, X_2, \ldots, X_d$, где d— число спектральных каналов.

Поскольку задача поиска варианта разбиения, оптимального по заданному критерию, имеет, как правило, экспоненциальную трудоемкость, на практике чаще всего применяются приближенные итеративные алгоритмы, которые на каждом шаге проводят модификацию текущего разбиения, дающую локальное улучшение качества.

При использовании коллективного подхода к кластерному анализу первоначально строится базовый набор вариантов группировки, по которым затем определяется итоговое разбиение на кластеры. Исходные решения формируются с использованием различных алгоритмов, по различным настройкам одного алгоритма, по случайно отобранным подсистемам переменных и т. п.

Существует несколько основных способов построения итоговых коллективных решений кластерного анализа. В первом способе от ансамбля требуют консенсуса, т. е. некоторой наилучшей степени согласованности с результатами отдельных алгоритмов.

Пусть имеется L вариантов $P_1, \ldots, P_L$ разбиения множества A на кластеры. Консенсусным разбиением называют такое разбиение P^*, для которого выполняется условие

$$P^* = \arg max_{P \in \mathcal{P}} \sum_{l=1}^{L} \phi(P, P_l),$$

где $\mathcal{P}$ — множество всевозможных разбиений A; ϕ — некоторая мера сходства между двумя разбиениями. В качестве меры сходства можно использовать индекс Ранда [14].

Пусть $P_1 = \{C_{1,1}, \ldots, C_{K1,1}\}$ и $P_2 = \{C_{1,2}, \ldots, C_{K2,2}\}$ — два варианта группировки; $C_{k,1} = \{a_{i1}, \ldots, a_{iNk,1}\}, C_{l,2} = \{a_{j1}, \ldots, a_{jNl,2}\}$, где $N_{k,1}$ — число объектов в k-м кластере первого варианта группировки, а $N_{l,2}$ — число объектов в l-м кластере второго варианта группировки. Индекс Ранда определяется как величина

$\varphi R(P_1, P_2) = \frac{(A+D)}{G}$, где A — число пар объектов, которые входят в одни и те же группы — в $P1$ и $P2$; D — число пар, которые входят в разные группы; $G = \binom{N}{2}$ — число всевозможных пар. Таким образом, данный индекс равен относительному числу правильно классифицированных (по принадлежности к

кластерам) пар объектов; он принадлежит интервалу от 0 до 1; значение 1 соответствует полному согласию между двумя разбиениями.

Второе направление в теории коллективного кластерного анализа основано на вычислении коассоциативной матрицы (матрицы смежности, coassociation matrix), определяющей, как часто пары объектов оказываются в одном и том же кластере в разных вариантах разбиения.

Усредненная коассоциативная матрица определяется как

$$H = \frac{1}{L}\sum_{l=1}^{L} H_l$$

Где H_l - коассоциативная матрица для l-го варианта разбиения. Элемент $h_l(i,j)$ этой матрицы равен нулю, если пара a_i и a_j $(i \neq j)$ объединена в одну группу; $h_l(i,j) = 1$, если данная пара разделена в l-м варианте разбиения по разным группам, $i,j = 1,\ldots,N$.

Элементы усредненной матрицы могут рассматриваться как аналоги попарных расстояний между объектами: чем больше значение элемента, тем чаще соответствующая пара была разнесена алгоритмами, входящими в ансамбль, в разные кластеры, т. е. тем более “непохожими” являются данные объекты. Для получения итогового согласованного разбиения может быть использован алгоритм кластерного анализа, обрабатывающий таблицы попарных расстояний, на вход которого подается полученная матрица.

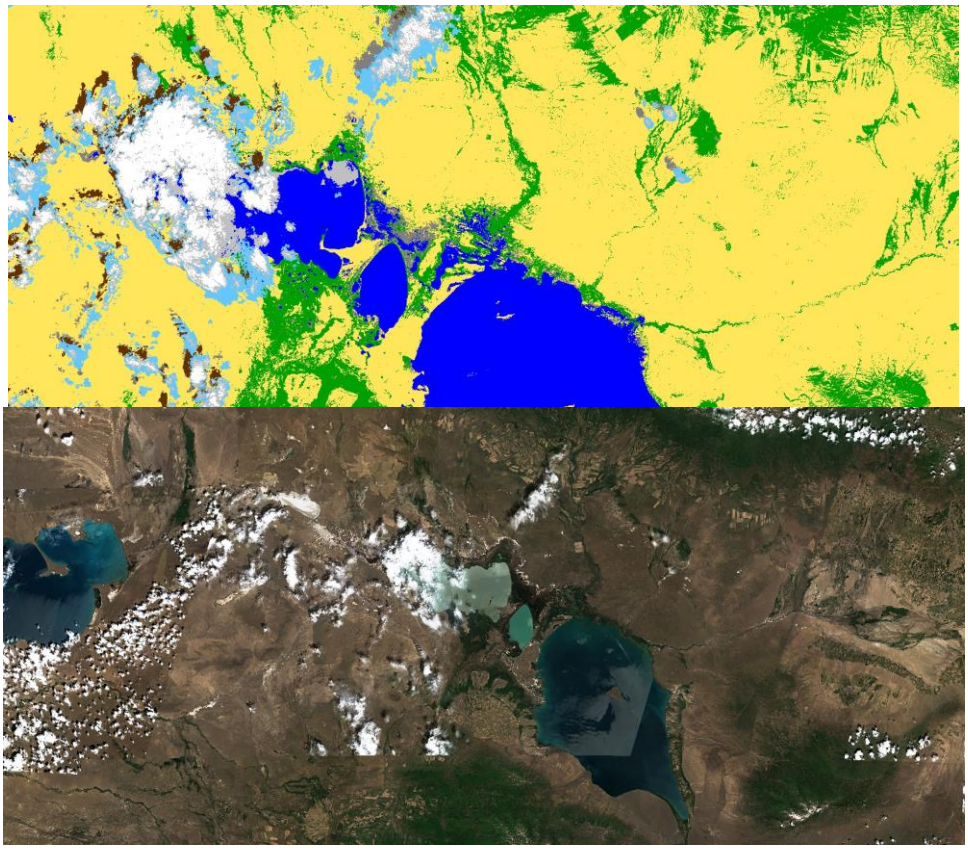
Метод построения ансамблевого решения на основе согласованной матрицы различий требует формирования и обработки матрицы размера N×N (N – число пикселей изображения), что существенно затрудняет его применение для сегментации изображений. Выходом из этой ситуации является переход от обработки отдельных элементов данных (пикселей) к обработке групп элементов. Способ формирования этих групп и выбора их представителей может зависеть от особенностей конкретного алгоритма.

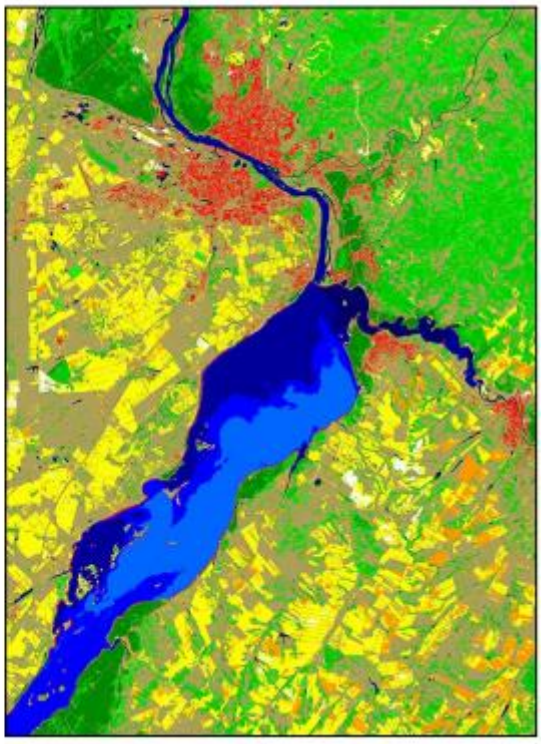
В данной работе предлагаются два способа формирования групп данных для построения ансамбля. Первый способ позволяет комбинировать результаты кластеризации, полученные произвольными алгоритмами. Элементы исходных данных объединяются в одну группу, если во всех разбиениях они отнесены в один кластер.

Согласованная матрица различий строится на множестве полученных групп. При этом подходе число групп пропорционально числу кластеров в разбиениях и мощности ансамбля. Второй способ позволяет формировать ансамблевое решение для непараметрических алгоритмов кластеризации, основанных на гистограммной оценке или оценке плотности Розенблатта–Парзена.

В этом случае каждый кластер содержит одну или несколько мод (локальных максимумов) плотности, которые определяются в процессе работы алгоритма. Данные разбиваются на группы, каждая из которых соответствует отдельной моде. Эти моды используются в качестве представителей отдельных групп. Согласованная матрица различий формируется на множестве.

В качестве базового используется разбиение, полученное при наименьшем значении параметра сглаживания (наиболее подробное разбиение из ансамбля). Группы относятся к тем же кластерам, что и их представители.





СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ИССЛЕДОВАНИЯ

Намоконов А. Н. Модели формирования набора данных для дистанционного зондирования Земли / А. Н. Намоконов, А.Л. Григорьева // молодежь и наука: актуальные проблемы фундаментальных и прикладных исследований материалы V Всероссийской. нац. науч. конф. студентов, аспирантов и молодых ученых, Комсомольск-на-Амуре, (Комсомольск-на-Амуре, апрель 2022 г.)..