Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное  
учреждение высшего образования  
«Самарский национальный исследовательский университет   
имени академика С.П. Королева»

УДК 004.932 На правах рукописи

Кравченко Кирилл Александрович

РАЗРАБОТКА ИНФОРМАЦИОННОЙ ТЕХНОЛОГИИ МНОГОМЕРНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА И КЛАСТЕРИЗАЦИИ ДАННЫХ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ

Автореферат

диссертации магистра

по направлению подготовки   
01.04.02 «Прикладная математика и информатика»

Магистерская программа   
«Высокопроизводительные вычисления и обработка данных»

Самара – 2016 год

Работа выполнена в Федеральном государственном автономном образовательном учреждении высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева» (Самарский университет).

Научный руководитель

*д.т.н., профессор В.В. Сергеев* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(*подпись*)

Рецензент

*к.т.н., доцент Н.И. Глумов* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(*подпись*)

Защита состоится « 09 » июня 2016 года на заседании Государственной экзаменационной комиссии по направлению 01.04.02 − «Прикладная математика и информатика» в Самарском университете, в ауд. 209 корпуса № 1, с 09-00 часов.

Секретарь Государственной экзаменационной

комиссии по направлению 01.04.02 −

«Прикладная математика и информатика»

д.ф.-м.н., доцент А.А. Ковалев \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(*подпись*)

# ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

## Актуальность темы

## При обработке многозональных аэрокосмических изображений, полученных от разных датчиков при дистанционном зондировании Земли, одной из важнейших процедур является сегментация (выделение типов земных покровов). В силу различных факторов (форма рельефа, изменчивость волнения и поля приводного ветра в районе акваторий, возраст льдов, наличие облачности» характер почв и растительности) регистрируемые приборами значения отраженной и собственной радиации носят случайный характер.

## Задача сегментации предполагает разбиение изображения на множество областей с некоторыми характеристиками однородности. Существуют различные подходы к определению понятия «однородности» выделяемых областей на изображении, рассматривающие текстурные признаки областей, геометрические характеристики и т.п.

## При отсутствии априорных сведений об изображении и объектах на нём целесообразно использовать поэлементную классификацию каждого его пикселя, что соответствует использованию в качестве характеристики «однородности» яркости пикселя изображения. Известные методы сегментации, реализующие такой подход, основаны на представлении изображений в виде смеси распределений, например, алгоритм EM (Expectation-maximization). Данный алгоритм при известном числе компонент смеси позволяет оценить параметры распределений и вклад каждой компоненты в смесь. Последующая сегментация производится с учётом максимальной апостериорной вероятности принадлежности пикселя конкретному распределению.

## Цель и задачи исследования

Цель настоящей работы – разработка информационной технологии, позволяющей максимально эффективно осуществить кластеризацию изображений с помощью модифицированного SHEM-алгоритма.

Для достижения цели исследования необходимо обеспечить решение следующих задач:

1. разработка структуры хранения данных об изображении;
2. осуществление программной реализации SHEM – алгоритма;
3. исследование возможностей модификации SHEM – алгоритма;
4. исследование результатов работы SHEM – алгоритма на тестовых изображениях.

## Научно-практическая новизна и значимость полученных результатов

Разработанную информационную технологию можно использовать для сегментации больших изображений с различным количеством слоев. Например, для снимков местности, полученных со спутника. Разработанное программное обеспечение позволяет быстро и дешево обрабатывать изображения больших размеров.

## Личный вклад магистранта

Использование дерева в качестве структуры хранения данных. Использование расстояния Бхаттачария для улучшения эффективности работы SHEM – алгоритма, главным недостатком которого является необходимость задавать количество распределений в качестве входного параметра.

## Апробация результатов магистерской диссертации

Основные результаты работы были представлены на LXVI Молодёжной научной конференции, посвященной 55-летию первого полёта человека в космос (5-9 апреля 2016, Самара, Россия); II международной конференции и молодёжной школе Информационные технологии и нанотехнологии (ИТНТ-2016) (17-19 мая 2016, Самара, Россия).

## Структура и объем диссертации

Диссертация состоит из введения, трех глав, заключения. Изложена на 60 страницах, содержит 27 рисунков, 10 таблиц. Список использованных источников составляет 20 наименований.

# ОСНОВНОЕ СОДЕРЖАНИЕ

**Во введении** дана общая характеристика работы, изложена актуальность, определена область исследований.

**Глава 1** содержит анализ задачи разделения смесей вероятностных распределений. Выведены формулы для работы базового ЕМ – алгоритма для случая, когда в смеси нормальные распределения. В данном случае плотность распределения имеет вид:

 где



Но при применении к реальным задачам (как например обработка спутниковых изображений) EM-алгоритма возникает ряд проблем. К таковым проблемам относятся, в первую очередь, медленная сходимость алгоритма и длительные вычисления. Чтобы улучшить данные аспекты, в данной работе используется модификация EM-алгоритма – SHEM-алгоритм («Statistical Histogram Based EM-algorithm»)]. Данная модификация отличается тем, что на каждой итерации мы обращаемся не к исходной выборке, а к ненормированной гистограмме, полученной из исходной выборки.

Введем множество . В данном множестве содержатся все неповторяющиеся значения, полученные из исходных данных. Также введем функцию  для обозначения количества элементов исходной выборки, соответствующих значению .

, где



Таким образом, в данном алгоритме формулы для вычисления параметров распределений будут модифицированы. Теперь суммирование будет происходить не по всем элементам входной выборки, а по всем элементам L, а значения конкретного элемента l будут домножаться на . В результате получим следующие формулы, определяющие работу алгоритма:









Также в данной главе описан способ хранения информации об изображениях.

В результате считывания изображения мы получаем многомерную выборку из N векторов вида

,

где N = a × b, a и b – длина и ширина изображения в пикселях соответственно,  – размерность вектора , .

Для случая, когда изображение представлено в цветовой схеме RGB мы получим , .

Для дальнейшей обработки этих данных нам необходимо выбрать структуру для их хранения. В данной работе мы используем модификацию EM-алгоритма, основанную на построении гистограмм (SHEM-алгоритм), поэтому для каждого возможного значения нам необходимо хранить количество векторов, имеющих данное значение. Наиболее удобно будет представить информацию в виде дерева, в котором корневой узел хранит в себе интервал, в который попадают все возможные пиксели изображения. Каждый из узлов данного дерева, за исключением узлов, отвечающих за конкретный пиксель, имеет потомков. Каждый из данных потомков делит интервал, на котором рассматривается частота встречи пикселей пополам по каждой координате. Пример первого уровня данного дерева для RGB модели можно увидеть на рисунке 1.

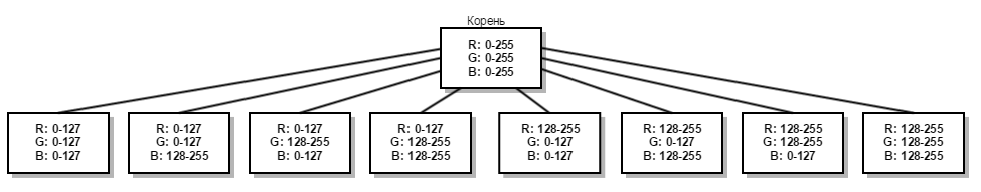


Рисунок 1 – Начало построения дерева для модели RGB

Было рассмотрено два способа построения дерева.

В первом варианте мы проходим по каждому пикселю изображения, и для полученных координат пикселя строим сразу всех необходимых потомков нашего дерева. Данный подход позволяет нам всего лишь один раз проходить по каждому пикселю изображения. На рисунке 2 изображен пример, как при RGB модели будет обрабатываться пиксель с координатами (154, 42, 251).

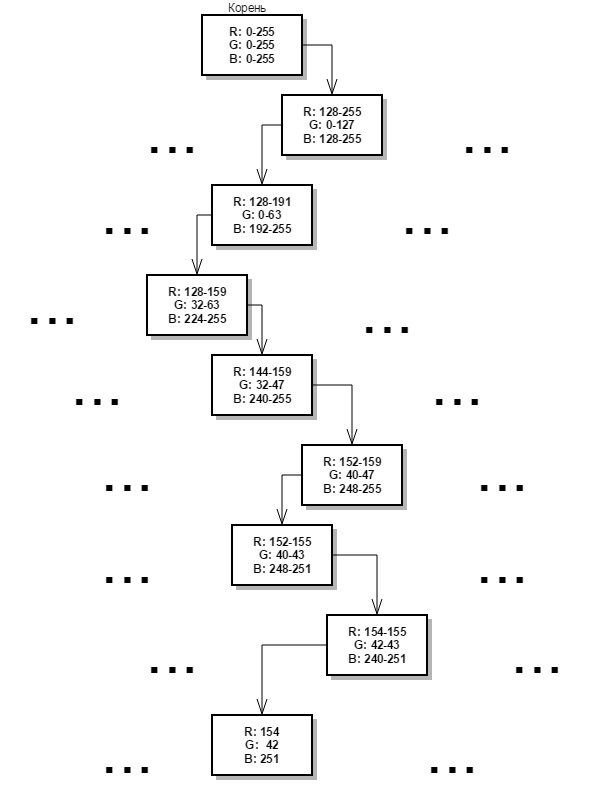


Рисунок 2 – Добавление RGB пикселя (154, 42, 251) в дерево

Во втором варианте будем считать, что корень дерева хранит в себе информацию обо всех пикселях. Пройдемся по всем пикселям, и для каждого из них определим потомка, в который попадет данный пиксель. Далее будем повторять эту процедуру для каждого элемента дерева, до тех пор, пока не дойдем до элементов, отвечающих за конкретный пиксель. При данном подходе по каждому из пикселей придется пройти несколько раз. Однако, для ускорения вычислений, представим значения каждого пикселя в двоичном виде. Результат в модели RGB для пикселя (154, 42, 251) можно увидеть на рисунке 3.

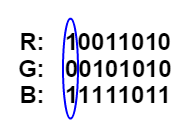


Рисунок 3 – Представление RGB пикселя (154, 42, 251) в двоичном виде

Когда будем определять потомка, в который нужно отправить данный пиксель в первый раз, возьмем значения первых цифр в двоичном представлении, как это показано на рисунке 1.3. В результате получим . Т.е. данный пиксель нужно отнести к пятому по счету потомку. Аналогично будем определять потомков и в дальнейшем, считывая необходимые значения цифр в двоичном представлении.

По результатам ряда экспериментов было получено, что для всех тестовых изображений второй вариант построения дерева работает быстрее. Таким образом, в дальнейшем будем использовать его для хранения данных об изображении.

**Глава 2** содержит описание разработанного программного комплекса.

Необходимость реализации параллельного алгоритма в рамках данной работы была продиктована большим объемом входных экспериментальных данных (изображения, полученные со спутников, могут иметь несколько десяткой слоев, которые нужно анализировать), а также, в случае многомерной модификации алгоритма SHEM, трудоемкостью вычислений.

Рассмотрим особенности реализации параллельного алгоритма. На E-шаге алгоритма находятся значения, определяющие вероятности того, что  значение в выборке принадлежит  компоненте в смеси. Вычисление данных значений можно распределить между процессорами по столбцам. В таком случае на каждой итерации должна быть выполнена нормализация значений в данных столбцах (чтобы избежать необходимости передачи данных между процессорами на данном шаге). На M-шаге алгоритма пересчитываются значения параметров всех компонент в смеси с помощью ранее полученных значений . Вычисление данных параметров распределяется между процессорами для отдельных распределений. Это требует передачи определенных данных между процессорами, например все полученные значения  должны быть доступны на каждом процессоре. Но, в нашем случае, когда рассматриваются только нормальные распределения, время коммуникации между процессорами не будет занимать много времени, если выполнить ряд предварительных вычислений для каждого из рассматриваемых распределений.

На рисунке 4 можно увидеть схему работы параллельного алгоритма.

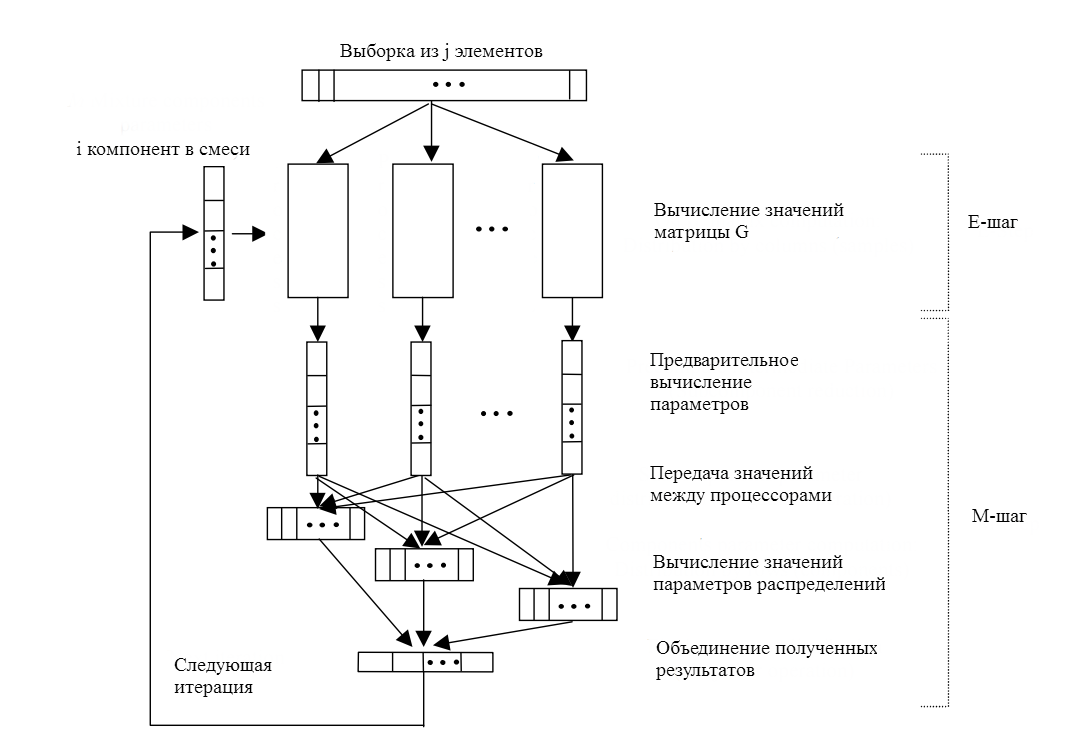


Рисунок 4 – схема работы параллельного алгоритма

В результате мы получаем, что время коммуникации между процессорами пренебрежимо мало по сравнению с непосредственным временем вычисления данных на каждом процессоре. Также коммуникации хорошо синхронизированы, то есть возникают одинаковое число раз для всех процессоров. Таким образом, каждый процессор не должен простаивать в ожидании результатов с других процессоров, и в то же время вычислительная нагрузка хорошо сбалансирована.

Проанализируем время работы параллельной программы, а также получаемое ускорение для разного количества процессоров. В таблице 1 показаны результаты работы программы над одним из изображений.

Таблица 1 – Время работы и ускорение программы для различного количества процессоров

|  |  |  |
| --- | --- | --- |
| Количество  процессоров | Время выполнения  программы (сек) | Ускорение |
| 1 | 54,535 | 1 |
| 2 | 28,812 | 1,892 |
| 4 | 15,109 | 3,609 |
| 8 | 7,588 | 7,187 |
| 16 | 3,974 | 13,722 |

На рисунке 5 представлен график зависимости времени работы программы от количества процессоров.

Рисунок 5

На рисунке 6 представлен график зависимости ускорения работы программы от количества процессоров.

Рисунок 6

Анализ полученных показал, что время выполнения программы ожидаемо уменьшается при большем числе процессоров, наибольшая эффективность параллелизма наблюдается при использовании восьми процессоров.

**В главе 3** приводятся результаты экспериментов по сегментации изображений с помощью разработанного программного комплекса.

Было проведено исследование зависимости времени работы программы от глубины построения дерева. Для эксперимента возьмем изображение, представленное на рисунке 6. Предположим, что на изображении пять распределений. Затем будем менять глубину строимого дерева, и сравним результаты. За наиболее точные данные примем глубину дерева, равную 8. В этом случае в дереве каждый элемент, не имеющий потомков, фактически отвечает за один конкретный цвет.



Рисунок 6

На рисунке 7 показан график зависимости времени выполнения программы от заданной глубины дерева.

Рисунок 7 – Зависимость времени выполнения программы от заданной глубины дерева

Из графика, изображенного на рисунке 7, видно, что с увеличением глубины дерева время выполнения программы растет экспоненциально. Но в то же время, если проанализировать результаты, более подробно представленные в работе, то видно, что полученные параметры распределений при больших глубинах не сильно отличаются. Поэтому был сделан вывод, что нет необходимости строить слишком подробное дерево. В рамках данной работы было решено для дальнейших экспериментов использовать глубину дерева, равную пяти.

Также было проведено исследование возможности уменьшения предполагаемого числа компонент в смеси. Алгоритм SHEM предполагает известное количество компонент смеси, что существенно ограничивает сферу его применения. Для преодоления этого ограничения предлагается использовать смесь распределений с заведомо избыточным количеством компонент с последующим слиянием тех из них, для которых пересечение плотностей вероятностей превосходит заданную величину.

В качестве дополнительной величины, определяющей необходимость слияния компонент можно использовать расстояние Бхаттачария между распределениями

Расстояние Бхаттачария – параметр, измеряющий сходство двух распределений. В случае многомерных нормальных распределений оно вычисляется по формуле:

,

где - вектора математических ожиданий,  - ковариационные матрицы, .

Для каждой пары из найденных распределений найдем данное расстояние. Чем оно меньше, тем более близки распределения друг к другу.

Запустим алгоритм на тестовом изображении, представленном на рисунке 6, поочередно при предположениях, что в смеси 3 – 10 распределений. В таблице 2 приведено минимальное расстояние Бхаттачария для этих случаев.

Таблица 2 – Минимальное расстояние Бхаттачария при различных предположениях о количестве распределений

|  |  |
| --- | --- |
| Количество распределений | Минимальное расстояние Бхаттачария |
| 3 | 3,1161 |
| 4 | 1,3727 |
| 5 | 1,4773 |
| 6 | 1,2947 |
| 7 | 1,0207 |
| 8 | 1,0451 |
| 9 | 0,9377 |
| 10 | 1,0954 |

Из данной таблицы видно, что при уменьшении предполагаемого количества распределений, расстояние Бхаттачария увеличивается, что и ожидалось. Таким образом, мы можем выбрать некий порог, и если минимальное расстояние его не превышает, то предположим, что распределений в смеси слишком много и уменьшим их количество на 1.

Например, если мы зададим такой порог, равный 1,2 для тестового изображения на рисунке 6, то это позволит при начальном предположении о 10 распределениях уменьшить их количество до 6.

# ЗАКЛЮЧЕНИЕ

В данной работе была выполнена поставленная цель, а именно был проанализирован EM – алгоритм, его модификация – SHEM–алгоритм, и их применение к задаче разделения смесей вероятностных распределений для обработки спутниковых изображений.

Была произведен анализ структур данных, подходящих для хранения информации об изображениях, и выбрана наиболее эффективная из них – дерево. Были рассмотрены способы построения деревьев, из которых был выбран наиболее быстрый.

Была разработана программная реализация SHEM-алгоритма, а также составлен ряд диаграмм, детально описывающих разработанную систему.

Основными направлениями исследований являлись результаты работы алгоритма при различных предположениях о количестве компонент в смеси, различной глубине дерева, а также различных входных изображениях.

В ходе исследования работы SHEM – алгоритма при различных параметрах были выявлены его недостатки. Основным недостатком является то, что алгоритм не позволяет определить количество компонент смеси. Нужно строить предположения о количестве компонент, и делать соответствующие выводы по полученным результатам. Было предложено использовать расстояние Бхаттачария для определения близости распределений, полученных в результате работы алгоритма. Анализ данного расстояния позволяет отсечь наиболее неправдоподобные предположения о количестве компонент в смеси.

Таким образом, ряд проведенных исследований позволяет оптимизировать работу SHEM-алгоритма: уменьшить время выполнения данного алгоритма, а также более точно предполагать количество элементов в имеющейся смеси распределений.

**СПИСОК ОПУБЛИКОВАННЫХ РАБОТ**

## Статьи:

1. Денисова А.Ю., Кравченко К.А., Сергеев В.В. Исследование алгоритмов оценки параметров смеси многомерных распределений. // II Международная Конференция и молодёжная школа Информационные технологии и нанотехнологии (ИТНТ-2016), 17-19 мая 2016, Самара, Россия, издательство СГАУ, 2016 г.