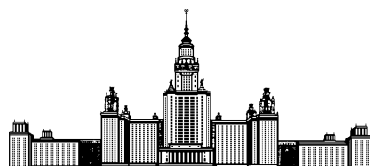


Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

**«Обработка астрономических снимков: многоуровневые
модели и эвристический подход»**

Выполнил:

студент 3 курса 317 группы

Славнов Константин Анатольевич

Научный руководитель:

д.ф-м.н., профессор

Дьяконов Александр Геннадьевич

Москва, 2014

Содержание

1	Введение	3
2	Trous вейвлет преобразование	5
3	Многоуровневое представление	8
3.1	Описание	8
3.2	Свойства	10
4	Значимые вейвлет коэффициенты	11
5	SExtractor	13
5.1	Оценка фона изображения	15
5.2	Фильтрация	15
5.3	Выделение	16
5.4	Объединение и разделение	16
5.5	Вычисление фотометрических данных	17
5.6	Морфологическая классификация галактик	17
6	Multiscale Vision Model	18
6.1	Определение Multiscale Vision Model	19
6.2	Поиск базовых структур	20
6.3	Граф межуровневой связности	21
6.4	Фильтрация	22
6.5	Объединение и разделение	22
6.6	Идентификация объекта	23
6.7	Задача частичного восстановления	23
6.8	Примеры	26
6.9	Выводы	28
7	Конкурс Galaxy Zoo	29
7.1	Специфика данных	29
7.2	Схема решения	29
7.2.1	Предобработка изображений	30

7.2.2	Извлечение признаков	33
7.2.3	Настройка метода	36
8	Благодарности	36

1 Введение

Астрономия — наука о Вселенной, изучающая расположение, движение, строение, происхождение и развитие небесных тел и образованных ими систем. В современном мире с появлением мощных телескопов и обсерваторий, количество данных, которые стали получать ученые выросло настолько, что обработка вручную становится невозможной. Поэтому необходимы алгоритмы и методы, которые бы справлялись со стоящими перед астрономами задачами. Данная проблема может быть решена с использованием техник анализа данных. Методы начали адаптировать для астрономических задач, изучать их поведение и улучшать. Данная курсовая работа обзорекает один из таких методов, применяемый в современной астрономии.

В основе своей пойдет речь про обработку изображений, а точнее фотоснимков небесных тел. Данные изображения имеют свои особенности. В самом плохом случае это нечеткий снимок отдаленного объекта достаточно низкого разрешения, зашумленный, возможно, с различными артефактами (например, черные полосы поперек объекта. Они возникают из-за того, что интересующее тело попало на стык матриц в телескопе). Помимо объекта интереса на снимке присутствует много лишних объектов, возможно часть из них накладывается на интересующий нас объект. И это далеко не все проблемы. На все перечисленные трудности нужно обращать внимание и учитывать их эффект.

Решение любой конкретной задачи анализа изображений в астрономии можно разбить на три этапа:

1. Сбор данных

Многие обсерватории предоставляют свои базы данных с тысячами изображений абсолютно свободно, которыми можно воспользоваться для оценки качества разных моделей и алгоритмов. Остается только сформировать из выбранного набора данных обучающую выборку. Однако, для получения верных ответов приходится привлекать экспертов.

Самое интересное, что и мнения экспертов могут отличаться в сложных ситуациях, т.е. на самом деле истинный ответ нам не будет известен, нам будет известна некая его оценка. Если величина ошибки алгоритма отличается от экс-

пертного мнения не больше, чем экспертные мнения отличаются между собой, то такой алгоритм можно назвать оптимальным по своей точности.

Интересным образом поступили в недавнем проекте Galaxy Zoo[2], цель которого была в том, чтобы собрать данные о морфологических характеристиках галактик по их изображению. Задача была решена с помощью краудсорсинга, т.е. множество людей отвечало на одни и те же вопросы про определенные свойства галактик. В итоге из всех ответов сформировали вероятности того, что галактика обладает некоторым свойством. Ответы хорошо совпадали с ответами, полученными от экспертов.

Таким образом, была получена хорошая классификация для многих галактик и появились данные, пригодные для машинного обучения. В дальнейшем был проведен конкурс на платформе kaggle.com с использованием этих данных[9].

2. Фильтрация и предобработка изображений

Важный этап в работе с изображениями небесных тел, которому стоит уделить отдельное внимание. Из изображения необходимо убрать все лишнее, выделить объект интереса, отмасштабировать, привести к нормальной форме. Это позволит избежать многих проблем на следующем этапе и увеличит качество получаемого результата.

3. Решение поставленной задачи

После того, как данные приведены в порядок, наступает основной этап работы. Большинство задач в астрономии являются задачами классификации, например, отделение галактик от звезд и других объектов, морфологическая классификация галактик, и другие.

Далее будет описан алгоритм, который называется Multiscale Vision Model, но перед тем, как переходить к непосредственному описанию, рассмотрим методы, которые лежат в его основе.

2 Trouis вейвлет преобразование

В данном разделе будет представлен один из типов вейвлет преобразования, который лежит в основе и потребуется в дальнейшем для описания Multiscale Vision Model.

Данный тип вейвлет преобразования дискретных сигналов известен как a trous (с дырками, такое название вызвано определенным методом свертки с пропусками, описанным во втором пункте алгоритма). Это преобразование без потери информации.

На выходе у данного метода J изображений одинакового размера, где J задано заранее. Размер совпадает с размером исходного изображения. Каждое изображение с номером j далее будет называться j уровнем и обозначаться w_j . Точку на позиции l j -ого уровня обозначим как $w_{j,l}$. Уровни упорядочены таким образом, что самый первый отвечает за высокочастотную составляющую изображения, а последний за низкочастотную.

Далее преобразование будет описано для одномерного сигнала, чтобы не загромождать формулы индексами. В двумерном случае все строится таким же образом.

Пусть есть некоторый низкочастотный дискретный фильтр h и исходное изображение c_0 . Каждый уровень вейвлет преобразования будет вычисляться как разность двух сигналов:

$$w_{j+1} = c_{j,l} - c_{j+1,l},$$

где сигнал c_j вычисляется как

$$c_{j,l} = \sum_k h(k)c_{j-1,l+2^j * k}.$$

Т.е. происходит некоторое подобие свертки предыдущего уровня с фильтром, только расстояние между соседними точками, которые участвуют в свертке, увеличивается как степень двойки. Принципиальная схема работы данного метода описывается рис. 1.

Полный алгоритм trous вейвлет преобразования выглядит так:

1. j инициализируется нулем, а $c_{0,l}$ начальными сигналом.
2. Вычисляем дискретную “свертку” $c_{j,l}$ с фильтром h . Расстояние между двумя соседними пикселями, участвующих в свертке равно 2^j .

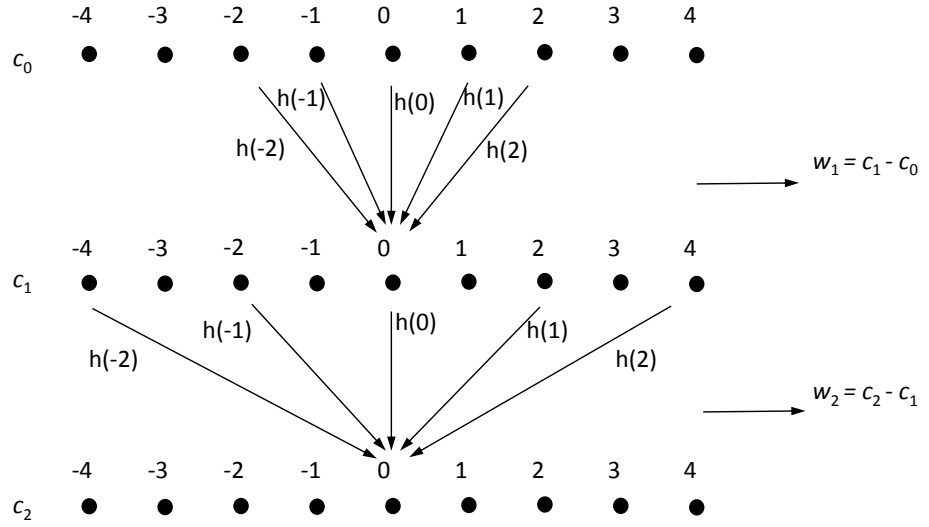


Рис. 1: Схема перехода от c_0 к c_1 и от c_1 к c_2 и получения 1 и 2 вейвлет уровней.

3. Вычисляем дискретное вейвлет преобразование на j уровне как разницу сигналов на текущей и последней выполненной итерацией:

$$w_{j+1,l} = c_{j,l} - c_{j+1,l}$$

4. Если j не достигло максимального заданного значения J , в плоть до которого мы хотим произвести расчет, мы увеличиваем j и переходим к шагу 2.
5. Множество $\mathcal{W} = w_1, w_2, \dots, w_J, c_J$ является вейвлет преобразованием исходного сигнала.

Такое преобразование позволяет быстро осуществить обратное: достаточно сложить все его компоненты:

$$c_{0,l} = c_{J,l} + \sum_{j=1}^J w_{j,l}$$

Отметим, что trous вейвлет преобразование можно назвать многоуровневым (multiresolution), т.к переводит одно изображение в набор из $J + 1$ изображений. При чем каждое из изображений отвечает за свой уровень, т.е. за свой масштаб

или разрешение исходного изображения. Смысл подобной терминологии наглядно пояснен на следующих примерах.

На рис. 2 приведен пример работы алгоритма над моделированным спектром, который изображен в верхней части рисунка. На рис. 4 приведен пример работы алгоритма в двумерном случае, примененный к галактике NGC 2997 на рис. 3

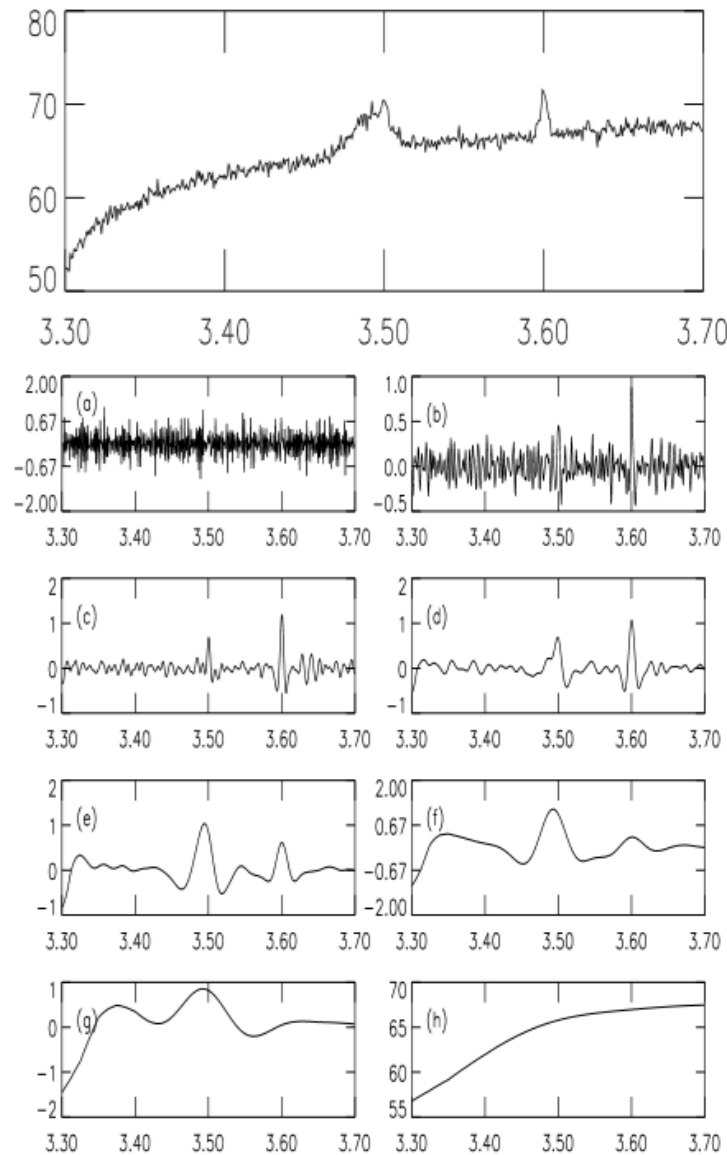


Рис. 2: На изображении вверху представлен смоделированный спектр. Далее сигналы (a-g) соответствуют различным уровням вейвлет преобразования с первого до 7 соответственно. сигнал (h) - есть остаточное сглаженное изображение c_J . Сумма всех восьми сигналов (a-h) в точности совпадет с исходным спектром.[1]

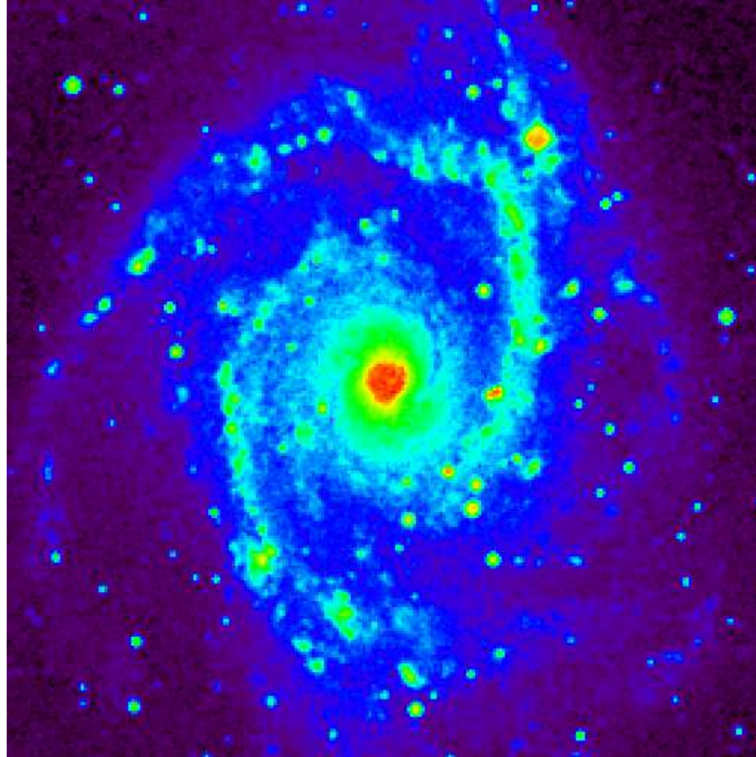


Рис. 3: Изображение галактики NGC 2997.[1]

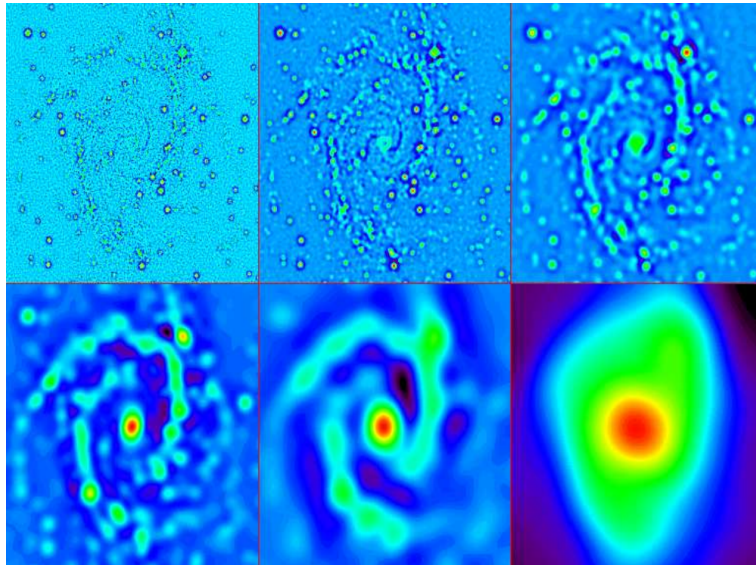


Рис. 4: Вейвлет преобразование галактики NGC 2997. нижнее правое изображение соответствует остаточному изображению C_J . [1]

3 Многоуровневое представление

3.1 Описание

Многоуровневое представление M (Multiresolution support) данных s — это бинарное представление данных, которое строится из их многоуровневого вида, например, из

описанного выше вейвлет преобразования. Если данные s содержат информацию на уровне j в позиции l , то $M_{j,l} = 1$.

M зависит от нескольких параметров:

1. Данные на входе.
2. Алгоритм, который используется для представления данных в виде многоуровневой структуры.
3. Шум в данных
4. Дополнительные накладываемые ограничения, которым должно соответствовать многоуровневое представление.

Многоуровневое представление данных может быть вычисленно следующим образом:

1. Вычисление вейвлет преобразования данных s (в нашем случае изображения), т.е. представление данных в виде многоуровневой структуры:

$$w = \mathcal{W}s = \{w_1, w_2, \dots, w_J, c_J\}.$$

2. Путем бинаризации каждого уровня приходим к многоуровневому представлению:

$$M = M_1, M_2, \dots, M_j, M_{j+1},$$

при чем на последнем уровне M_{j+1} все значения приравниваются единицам.

3. Можно скорректировать полученное представление, если имеются априорные знания.

Последний этап формируется из информации, которая есть о наших данных. Например, если мы знаем, что на нашем изображении нет объектов интереса больше или меньше определенного размера, то мы можем занулить все то, что относится к объектам такого типа. В таком случае удобно использовать методы математической морфологии.

На втором этапе необходимо вычислять какой из коэффициентов будет являться значимым, а какой нет, то есть для $j \leq J$ многоуровневое представление определено как:

$$M_{j,k} = \begin{cases} 1, & \text{если } w_{j,k} \text{ — значимый,} \\ 0, & \text{если } w_{j,k} \text{ — не значимый.} \end{cases}$$

Определение значимых коэффициентов будет описано в следующем разделе.

Для изображений можно визуализировать получившийся результат. Представим его в следующем виде:

$$I'_{k,l} = \sum_{j=1}^J 2^j M_{j,k,l}$$

На рис. 5 изображено представление галактики NGC 2997 в таком виде.

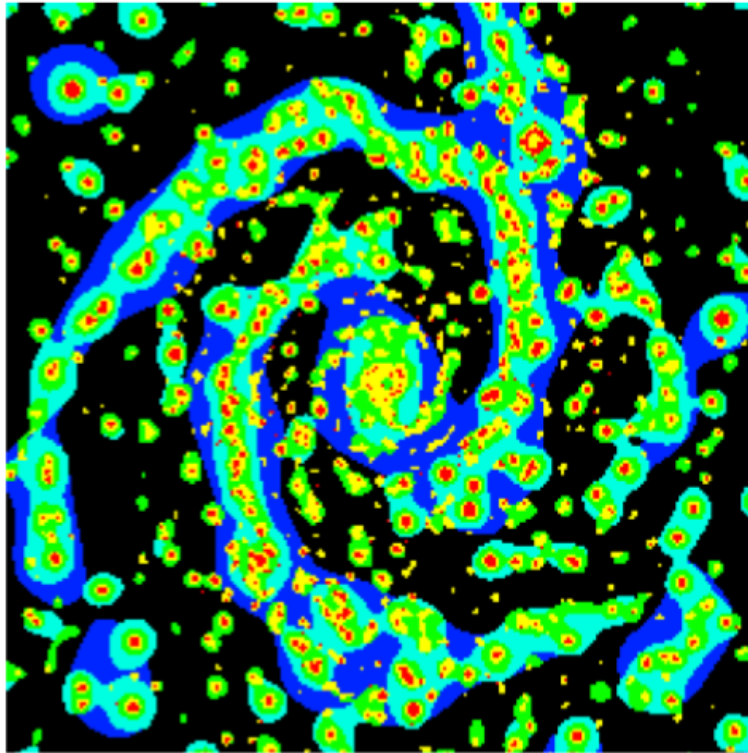


Рис. 5: Многоуровневое представление спиралевидной галактики.

3.2 Свойства

Самая важная часть преобразования заключена в его первом этапе, т.е. в trous вейвлет преобразовании. Все свойства trous вейвлет преобразования переносятся и на

многоуровневое представление. Поэтому далее будут разобраны свойства trous вейвлет преобразования.

В первую очередь такое преобразование — изотропное, что является преимуществом для изображений, объекты на которых не ориентированны. А астрономические тела как раз относятся к такому рода объектам.

У вейвлет преобразования есть 2 неудобных свойства: отрицательные значения коэффициентов и отсутствие устойчивости. По определению, среднее значение вейвлет коэффициентов — ноль. В процессе фильтрации данных могут возникать различного рода артефакты, которые усложняют анализ. Например, если мы отсекаем коэффициенты с маленькими значениями, и затем восстанавливаем изображение, общая яркость объектов на изображении изменится. Также в ситуациях, когда отрицательные коэффициенты имеют большой вес, подобная операция может привести к неправильному определению структур на изображении.

Неустойчивость оказывает сильное влияние на результат тем, что точечные артефакты, которые неизбежно возникают при астросъемке, “засоряют” все уровни вейвлет преобразования, что ведет к ухудшению последующей обработки.

Единственный способ борьбы с проблемами данного метода оставить основную логику многоуровневого преобразования, основанного на trous вейвлет преобразовании, но использовать другие, не вейвлет функции. Например, использовать медианный фильтр.

Вообще вейвлет преобразование не является обязательным. В зависимости от данных, можно использовать и другие многоуровневые преобразования. Например, если изображения содержат много вытянутых объектов, ридгелет(ridgelet) или курвлет(curvelet) преобразования могут показать результат лучше, чем вейвлет преобразование.

4 Значимые вейвлет коэффициенты

Обычно изображения содержат шум. Значит и вейвлет коэффициенты тоже зашумлены. В большинстве приложений необходимо знать, относятся ли вейвлет коэффициенты к сигналу(т.е. являются значимыми) или к шуму.

Вейвлет преобразование на уровне j описывается вейвлет коэффициентами $w_{j,k}$. Если мы получим распределение коэффициентов $w_{j,k}$ на каждом из уровней, то можно применить тест статистической значимости. Пусть \mathcal{H}_0 — гипотеза, что изображение является локально постоянным на уровне j . Отказ от гипотезы \mathcal{H}_0 зависит от величины

$$P = \Pr(|w_{j,k}| < \tau | \mathcal{H}_0).$$

Порог чувствительности τ определяется для каждого уровня в отдельности. Если $P = P(\tau) > \epsilon$, то гипотеза \mathcal{H}_0 не опровергается. Если $P < \epsilon$, то значение коэффициента не может быть связано только с шумом, поэтому гипотеза \mathcal{H}_0 опровергается. В таком случае был обнаружен значимый коэффициент (см. рис 6)

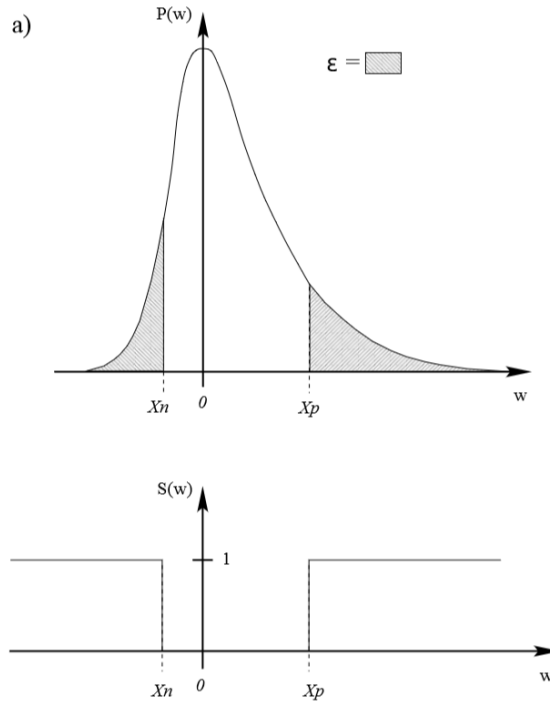


Рис. 6: Поиск значимых коэффициентов. Если значение вероятности меньше определенного порога ϵ , то коэффициент считается значимым. [1]

Отметим, что таким же образом изображение может подвергаться фильтрации. Достаточно только занулять незначительные коэффициенты, оставляя значительные. Но в большинстве методов фильтрации, основанных на PCA или преобразовании Фурье, объект восстанавливают используя или самые важные его компоненты в

случае PCA или по определенной частотной полосе в случае преобразования Фурье. В нашей терминологии можно сказать, что изображение восстанавливается не по всем уровням. Отличие приведенного метода фильтрации как раз в том, что фильтрация происходит независимо на всех уровнях, после чего все уровни участвуют в восстановлении изображения.

5 SExtractor

В этом разделе будет приведено описание общего подхода к обработке астрофотоснимков с целью сегментации изображения и извлечения информации о изображенных объектах, на котором основана Multiscale Vision Model.

SExtractor(Source Extractor)[3] — программное обеспечение, которое автоматически выделяет, отделяет друг от друга и классифицирует объекты на снимках космоса. Как пишут авторы, метод, положенный в основу программы позволяет работать с очень большими изображениями(до 60 000 x 60 000 пикселей), позволяет осуществлять надежное разделение галактик и звезд.

Данная программа примечательна тем, что в ней заложен интересный подход к выделению объектов на астрономических изображениях. Схема состоит из следующих шагов:

1. Оценка фона изображения
2. Фильтрация
3. Выделение объектов
4. Объединение и разделение объектов
5. Вычисление фотометрических данных(техника, применяемая в астрономии для измерения потока или интенсивности электромагнитного излучения астрономического объекта)
6. Классификация

Далее будет подробнее описан каждый шаг.

Обычно астрономические изображения содержат больше количество точечных источников света(звезды), квази-точечные источники(далекие нечеткие галактики, двойные звезды) и сложные рассеянные структуры(галактики, туманности, кластеры). Часто объекты на снимке организованы иерархично, например, звезда лежит поверх небольшой туманности, которая лежит в рукаве галактики, а сам рукав является частью галактики.

Обычно рассеянные структуры типа галактики менее яркие относительно точечных источников света. Поэтому если попытаться выделять объекты по яркостному порогу, большинство рассеянных структур не будут найдены. Если же сделать порог меньше, появится большое количество шума, качество выделения сильно пострадает(см. рис. 7). Подобная ситуация мотивировала на создание более тонкого алгоритма, который получил название Multiscale Vision Model(MVM). Под Vision Model подразумевается последовательность операций, необходимых для автоматической обработки изображения.

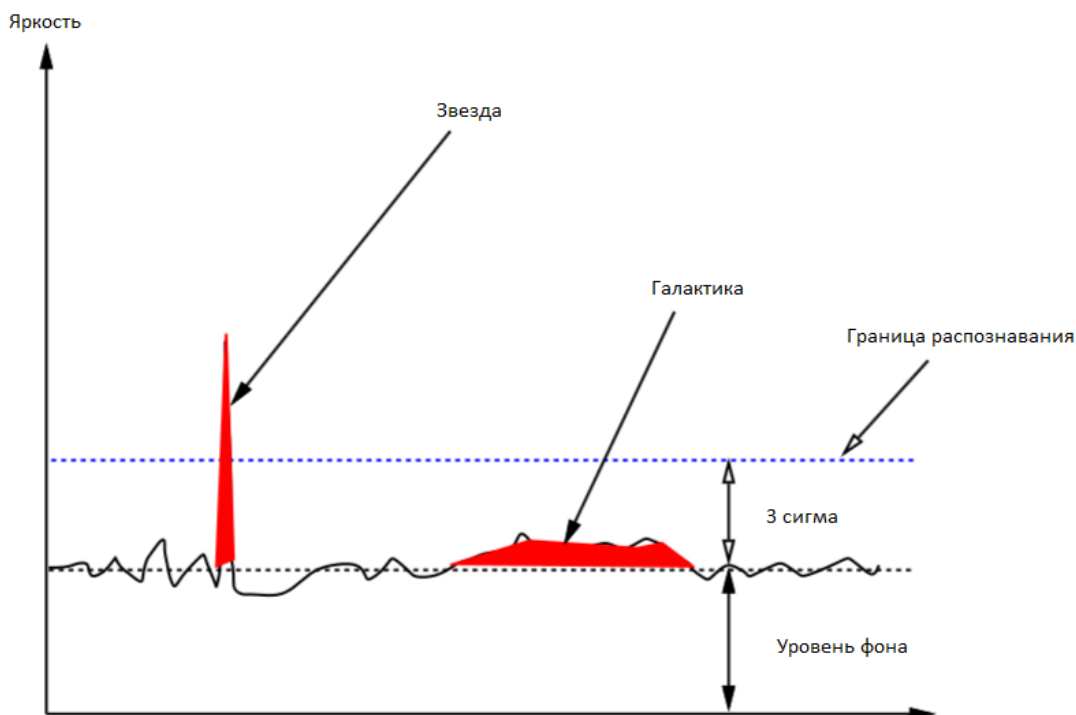


Рис. 7: Пример данных, на которых изображен рассеянный и точечный источник света. рассеянный объект может быть выделен глазом, но стандартные методы его не находят.[1]

5.1 Оценка фона изображения

Обычно изображение интересующего нас объекта смешивается с плоским сигналом, который называется фоновым. Фоновый сигнал нужно точно оценить. Если его оценить не точно, то оценки яркости объектов будут смещены.

Предлагается следующий способ: изображение делится на блоки и по гистограмме оценивается локальный уровень фона в каждом блоке. $p(I)$ моделируется используя 3 величины: настоящий уровень фона s , нормальный шум с дисперсией σ и параметр a определяющего неравномерность $p(I)$, которая возникает из-за объектов, расположенных поверх фона:

$$p(I) = \frac{1}{a} \exp\left(\frac{\sigma^2}{2a^2}\right) \exp\left(-\frac{I-s}{a}\right) \operatorname{erfcf}\left(\frac{\sigma}{a} - \frac{I-s}{\sigma}\right)$$

Применение медианной фильтрации к получившейся оценки фона поможет исправить неправильные значения, например, в тех местах, где большую часть блока занял яркий объект. Применяя билинейную или бикубическую интерполяцию получим итоговую оценку.

В данном методе важно хорошо подобрать размер блока. Если он окажется очень маленьким, в фон попадут и некоторые объекты, если слишком большим, метод не учтет истинную вариативность фона.

Однако, это не единственный метод оценки фона. например, в [4] оценку фона в блоке производят с помощью 3-сигма отсечения вокруг медианы. Если при отсечении стандартное отклонение изменилось меньше чем на 20%, то блок считается отчищенным от не фоновых пикселей, и уровень фона полагается равным среднему по гистограмме. Иначе он приравняется $c_1 \times \text{median} - c_2 \times \text{mean}$, где c_1, c_2 некоторые константы, которые могут варьироваться. например $c_1 = 2.5, c_2 = 1.5$. Этот подход предпочтительнее, чем первый, т.к. он быстрее с вычислительной точки зрения и более устойчив при малом размере изображения.

5.2 Фильтрация

Для качественного выделения изображение должно быть отфильтровано. При чем форма фильтра оптимизирует поиск объектов с такой же формой. Например, для вы-

деления звезд оптимальным фильтром является PSF. Для поиска рассеяных структур лучше выбрать больший размер фильтра.

Если же необходимо выделить объекты любой формы, их поиск необходимо повторить несколько раз с разными размерами фильтров, что приводит к подобию многоуровневого (multiscale) подхода.

5.3 Выделение

После того, как изображение отфильтровано, все пиксели I_l в позиции l , значение которых больше, чем $T_l = B_l + k\sigma$ считаются значимыми. B_l — оценка фона изображения, σ — стандартное отклонение шума и k — заданный коэффициент, обычно между 3 и 5. Таким образом, путем зануления незначимых пикселей получаем изображение, разделившееся на сегменты. Поставим в соответствие каждой связанной компоненте свой уникальный номер. На следующем этапе будет произведено разделение смешавшихся объектов, которые имеют одинаковый номер.

Альтернативный способ выделения объектов — поиск пиков яркости. Данный метод хорошо подходит только для выделения звезд, но не рассеяных объектов. В таком случае, на следующем этапе необходимо объединять компоненты, которые на самом деле принадлежат к одному объекту.

5.4 Объединение и разделение

Самый важный этап. Рассеяные объекты должны считаться одним объектом, в то время как остальные тела должны быть отделены друг от друга. В SExtractor каждая группа связанных пикселей анализируется на различных уровнях яркости. Рассматривая изображение как поверхность, самом высоком уровне яркости видны только самые высокие пики. Далее с убыванием яркости уровня, появляются некоторые другие пики, определяя новые объекты. На определенном уровне, 2 объекта могут объединиться. Тогда приходится решать, являются ли они одним объектом или же их необходимо разделить. Такое решение может быть осуществлено на основе информации о яркости каждого из пиков: если разница в яркости невелика, то объекты объединяются в один.

5.5 Вычисление фотометрических данных

Достаточно важная задача в астромии, которая позволяет установить многие свойства небесных тел. Фотометрический анализ позволяет получить информацию о распределении массы в галактиках, об их глобальной структуре и геометрических параметрах. Многоцветная фотометрия позволяет сделать заключения о звездном населении галактик, об их пространственной ориентации, о наличии и характеристиках пылевой составляющей и т.д.

Существуют методы, которые с достаточно хорошей точностью определяют фотометрию объекта. Основной задачей в таких методах является поиск области, по которой будет вестись дальнейшие вычисления. Адаптивный метод (Adaptive aperture photometry) использует первый момент изображения, чтобы определить область эллиптической формы, по которой будет просуммирована яркость объекта. Подробнее материал изложен в [5].

5.6 Морфологическая классификация галактик

В связи со сложностями, которые сопряжены с описанием спиральных галактик, особенно если они развернуты к нам боком, большинство работ концентрируется на эллиптических свойствах галактик. Но и здесь возникает много проблем. Поэтому большинство работ не решают данную проблему в прямую, а делают выводы о качестве признаков, которые они использовали.

Если использовать подход описанный выше, т.е. применять мультимасштабные статистические алгоритмы, то и признаки должны быть однородной природы. Выделяется два основных подхода к выбору признаков:

1. Выбор профиля яркости вдоль основных осей объекта. Зашумленность и нечеткость таких признаков требует большого внимания к надежности выделения. Такой профиль может быть определен достаточно надежно, если предположить, что галактика развернута к наблюдателю, является плоской, симметричной вдоль основных осей.
2. Таких проблем не возникает если работать с морфологическими признаками.

Яркость галактики можно попытаться приблизить непрерывной функцией, а затем из знания этой функции выделить новые признаки. Существуют законы, которые описывают как яркость точки зависит от расстояния от центра галактики. Например, она может убывать как $r^{\frac{1}{4}}$ или по экспоненциальному закону. Существует закон Серсика — достаточно общий эмпирический закон распределения поверхностной яркости галактики:

$$\ln \frac{I(R)}{I_0} = -kR^{1/n},$$

Где I_0 — яркость при $R = 0$, n — индекс Серсика, регулирующий плавность убывания яркости: при большом значении ($n > 4$) яркость убывает почти линейно. Если n около единицы, то в определенный момент значение яркости начинает резко убывать, до этого моменты являлась почти константной.

Однако, подобные законы подразумевают однородность галактики. Если же пытаться обнаруживать спиралевидные элементы галактик, необходим иной подход. Существуют подходы, которые основываются на разбиении изображения на восемь угловых секторов. Так же преобразование Фурье могут выявлять спиральную структуру. [7]

Далее полученные признаки могут обрабатываться используя стандартные методы кластеризации и классификации, PCA, дискриминантного анализа и т.д.

6 Multiscale Vision Model

Мультимасштабное(многоуровневое) преобразование изображения — это преобразование, при котором определенный алгоритм обработки применяется к каждой j -ой шкале из $\{w_j\}$, где $\{w_j\}$ является многоуровневым представлением изображения. На выходе получается изображение такого же размера как и исходное. Например, trous вейвлет преобразование является многоуровневым, и исходное изображение I может быть представлено в виде суммы всех вейвлет шкал и низкочастотного изображения-остатка c_J :

$$I = c_J + \sum_{j=1}^J w_j.$$

Если теперь определенный алгоритм \mathcal{A} применить к каждому уровню

$$w'_j = \mathcal{A}(w_j), \text{ для } j = 1..J,$$

$$c'_J = \mathcal{A}(c_J),$$

Получится новое изображение:

$$I' = c'_J + \sum_{j=1}^J w'_j.$$

Что и является многоуровневым преобразованием изображения.

На следующем этапе нам необходимо описать уже не изображение, а интересующий нас объект в многоуровневом виде. В зависимости от типа объектов, которые мы ходим разделять, представления будут разные. В случае астрономических тел можно поступить следующим образом, разложить изображение на набор компонент:

$$I(k, l) = \sum_{i=1}^{N_0} O_i(k, l) + B(k, l) + N(k, l),$$

где N_0 — количество объектов, O_i — объекты, содержащиеся в данных (звезды, галактики и т.д.), B — фон, N — шум.

Для того, чтобы реализовать такое разложение, необходимо проделать шаги, описанные в предыдущем пункте. По сути Multiscale Vision Model является реализацией той схемы и предоставляет алгоритм действий. Основная идея его работы заключается в определении связности отдельных компонент изображения. Вейвлет пространство двумерного изображения является трехмерным. Понятие связности в исходном пространстве основано на смежности пикселей друг с другом. Его можно перенести в вейвлет пространство. Далее будет подробно разобран данный метод.

6.1 Определение Multiscale Vision Model

Multiscale Vision Model (MVM) — описывает объект как иерархический набор структур [8]. В [8] используются следующие определения:

- Значимые вейвлет коэффициенты — если абсолютное значение коэффициента больше некоторого порога, то он считается значимым. Этот порог зависит от характера шума модели.

- Структура — это набор смежных значимых вейвлет коэффициентов на j -ой шкале:

$$S_{j,k} = \{w_{j,x_1,y_1}, w_{j,x_2,y_2}, \dots, w_{j,x_p,y_p}\},$$

где w_{j,x_i,y_i} — вейвлет коэффициент в j -ой шкале на позиции (x_i, y_i) , k — порядковый номер структуры.

- Объект — набор структур:

$$O_l = \{S_{j_1,k_1}, \dots, S_{j_n,k_n}\}.$$

Также определим оператор \mathcal{L} , который будет показывать к какому объекту принадлежит данная структура:

$$\mathcal{L}(S_{j,k}) = l, \text{ если } S_{j,k} \in O_l$$

$\mathcal{L}(S_{j,k}) = 0$, если структура не принадлежит ни одному объекту.

- Уровень объекта — равняется уровню вейвлет коэффициента с максимальным значением, который относится к объекту
- Межмасштабные(межуровневые) отношения — критерий, позволяющий объединять две структуры в одну.
- Подобъект — часть объекта. Появляется, если объект имеет локальный максимум в вейвлет пространстве. Подобная иерархия позволяет анализировать как объект, так и его часть.

6.2 Поиск базовых структур

После того как изображение представлено в многоуровневом виде, мы имеем бинарное изображение на каждом из уровней: пиксель имеет значение 1, если он стоит на позиции значимого коэффициента, иначе 0. На каждом уровне происходит поиск связанных компонент, которые нумеруются от 1 до L_{max} . Положим структуру $S_{j,i}$ равную компоненте связности с номером i на уровне j .

6.3 Граф межуровневой связности

Объект описывается как набор структур, которые объединяются по правилам межуровневых отношений. Определим эти правила. рассмотрим 2 структуры на двух соседних уровнях $S_{j,k}$ и $S_{j+1,k'}$. Каждая из этих структур занимает определенное положение в исходном пространстве, там где сигнал является значимым. Обозначим (x_m, y_m) как позицию пикселя, который имеет наибольшее значение в структуре $S_{j,k}$. Тогда $S_{j,k}$ будет объединена с $S_{j+1,k'}$, если $S_{j+1,k'}$ содержит пиксель с позицией (x_m, y_m) (то есть пиксель с максимальным значением в структуре $S_{j,k}$ должен содержаться и в структуре $S_{j+1,k'}$) (рис. 8). Объединяя или, можно сказать, соединяя всевозможные структуры по приведенному правилу, получим граф межуровневой связности. Набор вершин любой связной компоненты графа будет являться объектом по введенному выше определению.

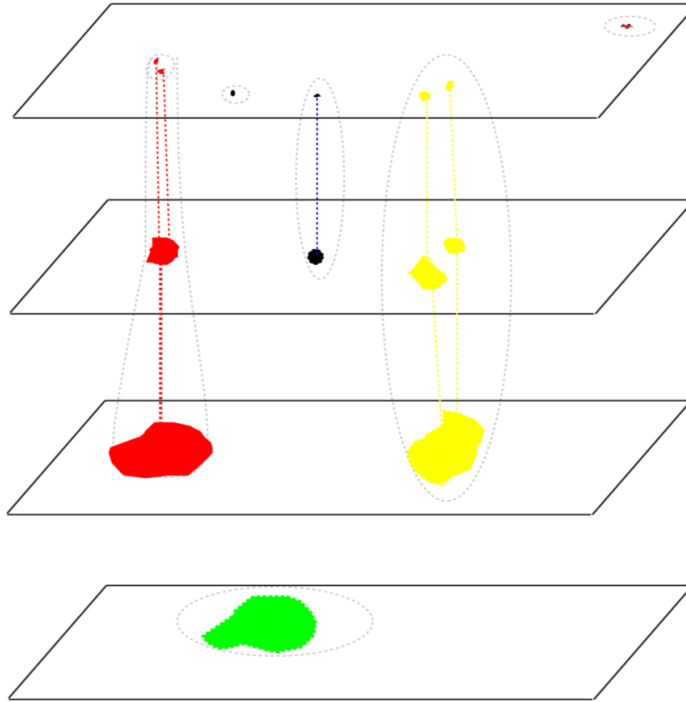


Рис. 8: Визуализация работы процедуры по поиску объектов на снимке, пример связности в вейвлет пространстве: Две структуры на двух соседних уровнях S_j, S_{j+1} связаны тогда, когда пиксель с максимальным вейвлет коэффициентом в структуре $S_{j,k}$ содержится и в структуре $S_{j+1,k'}$. [1]

6.4 Фильтрация

Некоторые структуры могут образоваться из-за зашумленности данных. Большинство из них содержит всего несколько изолированных (не связанных ни с какими структурами на других уровнях) пикселей. Для того чтобы избежать обнаружения ложных объектов, подобные структуры должны быть удалены из графа межуровневой связности.

6.5 Объединение и разделение

Объекты на изображении могут быть сложно организованны и могут возникнуть ситуации, когда два близкорасположенных, но разных объекта слиплись в один. Рассмотрим две ситуации:

- Два (или более) близких объекта примерно одинакового размера создают набор структур. На некотором уровне j две различные структуры $S_{j,1}$ и $S_{j,2}$, а на $j+1$ уровне эти структуры объединяются в один объект.
- Два (или более) близких объекта разного размера создают набор структур с j -го по k -й уровень.

В вейвлет пространстве решение об объединении или разделении структур будет основано на значениях локального максимума различных структур данного объекта. Структура $S_{j,k}$ отделяется в новый объект, если существует хотя бы одна структура исходного объекта на этом же уровне, и выполняются соотношения:

$$w_j^m > w_{j-1}^m$$

$$w_j^m > w_{j+1}^m$$

где:

- w_j^m максимальный вейвлет коэффициент структуры $S_{j,k}$: $w_j^m = \max(S_{j,k})$
- $w_{j-1}^m = 0$ если $S_{j,k}$ не связан ни с какой структурой на $j-1$ уровне. Либо это максимальное значение вейвлет коэффициента структуры $S_{j-1,k'}$, где $S_{j-1,k'}$ такая что $\mathcal{L}(S_{j-1,k'}) = \mathcal{L}(S_{j,k})$ и позиция наибольшего вейвлет коэффициента является ближайшей к позиции максимума $S_{j,k}$

- $w_{j+1}^m = \max(w_{j+1,x_1,y_1}, \dots, w_{j+1,x_n,y_n})$, где $w_{j+1,x,y}$ вейвлет коэффициенты на позициях, которые также относятся к $S_{j,k}$.

Если эти условия выполнены, то структура, и все связанные с данной структуры на меньших уровнях определяют новый объект.

6.6 Идентификация объекта

Теперь можно привести полное описание данного метода, который позволяет нам найти все объекты на изображении I :

1. Вычисление trous вейвлет преобразования, результатом которого будет набор $W = \mathcal{W}I = \{w_1, \dots, w_J, c_J\}$. Каждый уровень w_j имеет такой же размер как и исходное изображение.
2. Вычисление стандартного отклонения шума на w_1 уровне
3. Вычисление порогов для каждого уровня исходя из модели шума
4. Бинаризация всех уровней по порогу и разделение изображения на связные компоненты
5. Определение межуровневых отношений
6. Определение максимумов в пространстве вейвлет коэффициентов.
7. Выделение всех связанных деревьев из графа межуровневой связности с вершинами в тех структурах, к которым относятся определенные на предыдущем пункте максимумы

6.7 Задача частичного восстановления

Частичное восстановление как обратная проблема. Обозначим набор структур, который определяет объект O_i как $S_i = \{S_{j,k}, \dots, S_{j',k'}\}$. Любой из объектов может быть восстановлен отдельно от остальных, а путем объединения всех объектов на одном изображении получим фильтрованное изображение. Обозначим W_i как набор вейвлет коэффициентов, относящихся к объекту O_i . С другой стороны, W_i это

часть вейвлет преобразования O_i : $\widetilde{W}_i = \mathcal{W}O_i$. Но часть коэффициентов \widetilde{W}_i неизвестна, полностью неизвестен весь последний уровень \widetilde{W}_i . Тогда задача восстановления может быть сведена к задаче поиска такого изображения O_i , что в его вейвлет преобразовании соответствующие коэффициенты были равны W_i . Если \mathcal{W} это оператор вейвлет преобразования, а P_w проекционный оператор в подпространство известных коэффициентов, то решение задачи может быть найдено с помощью минимизации функционала [8]:

$$J(O_i) = \|W_i - A(O_i)\|,$$

где оператор A определен как $A = P_w \circ \mathcal{W}$.

Перед нами же стоит обратная проблема: по известным A и W_i найти O_i . Решение задачи зависит от регулярности A . Размер восстанавливаемого изображения O_i может быть гораздо больше, чем количество известных коэффициентов. И существует только по крайней мере одно изображение(исходное), которое выдает именно коэффициенты W_i . Но в большинстве случаев мы имеем бесконечное множество решений, среди которых необходимо выбрать одно. Изображение это функция, значения которой всегда больше или равно нулю, что дает нам дополнительные ограничения, однако они не позволяют выделить единственное решение.

Алгоритм восстановления. Для решения задачи(найти такое O_i , такое что $W_i = A(O_i)$) можно применить метод наименьших квадратов:

$$\|W_i - A(O_i)\| \rightarrow \min$$

O_i является минимумом данного функционала тогда и только тогда, когда

$$\widetilde{A}(W_i) = (\widetilde{A} \circ A)(O_i),$$

где \widetilde{A} определено в [8] как:

$$\widetilde{A}(W_i) = \sum_{j=1}^J (h_1 * \dots * h_{j-1})W_i(j),$$

где h — это низкочастотный фильтр используемый при вейвлет разложении, а $W_i(j)$ — та часть коэффициентов W_i , которая относится к уровню j .

Алгоритм восстановления:

1. Инициализация: оценка изображения O_i^n , остаточный вейвлет w_r^n и остаточное изображение R^n инициализируются.

$$O_i^0 = \mathcal{W}^{-1}W_i, w_r^0 = W_i - A(O_i^0), R^0 = \tilde{A}(w_r^0),$$

где \mathcal{W}^{-1} - оператор обратного вейвлет преобразования. Изображение O_i восстанавливается из вейвлет коэффициентов W_i как сумма всех вейвлет уровней. Из-за того, что W_i не обязательно является вейвлет-преобразованием изображения O_i , то $\mathcal{W}^{-1}\mathcal{W}O_i$ может не совпадать с O_i .

2. Вычисление параметра близости α^n :

$$\alpha^n = \frac{\|\tilde{A}(w_r^n)\|^2}{\|\tilde{A}(R^n)\|^2}$$

3. Подсчет O_i^{n+1} :

$$O_i^{n+1} = O_i^n + \alpha^n R^n$$

4. Обнуление тех значений в O_i^{n+1} , которые оказались меньше нуля.

5. Вычисление остаточного вейвлета:

$$w_r^{n+1} = W_i - A(O_i^{n+1})$$

6. Если значение $\|w_r^{n+1}\|$ меньше, чем заданный порог, то выход из итерационного процесса.

7. Вычисление промежуточного параметра β^{n+1} :

$$\beta^{n+1} = \frac{\|\tilde{A}(w_r^{n+1})\|^2}{\|\tilde{A}(w_r^n)\|^2}$$

8. Вычисление остаточного изображения:

$$R^{n+1} = \tilde{A}(w_r^{n+1}) + \beta^{n+1}R^n$$

9. Вернуться ко второму шагу.

6.8 Примеры

Данный метод обработки изображений является мощным инструментом, который позволяет решить далеко не одну задачу. Например, задача фильтрации изображения, разделение изображений, наложенных друг на друга, разделение объекта на его структурные составляющие. На рис. 9 в верхней части изображен моделированный зашумленный неоднородным шумом спектр излучения некоторого объекта. После применения данного метода удалось точно выделить основную компоненту спектра и убрать весь шум

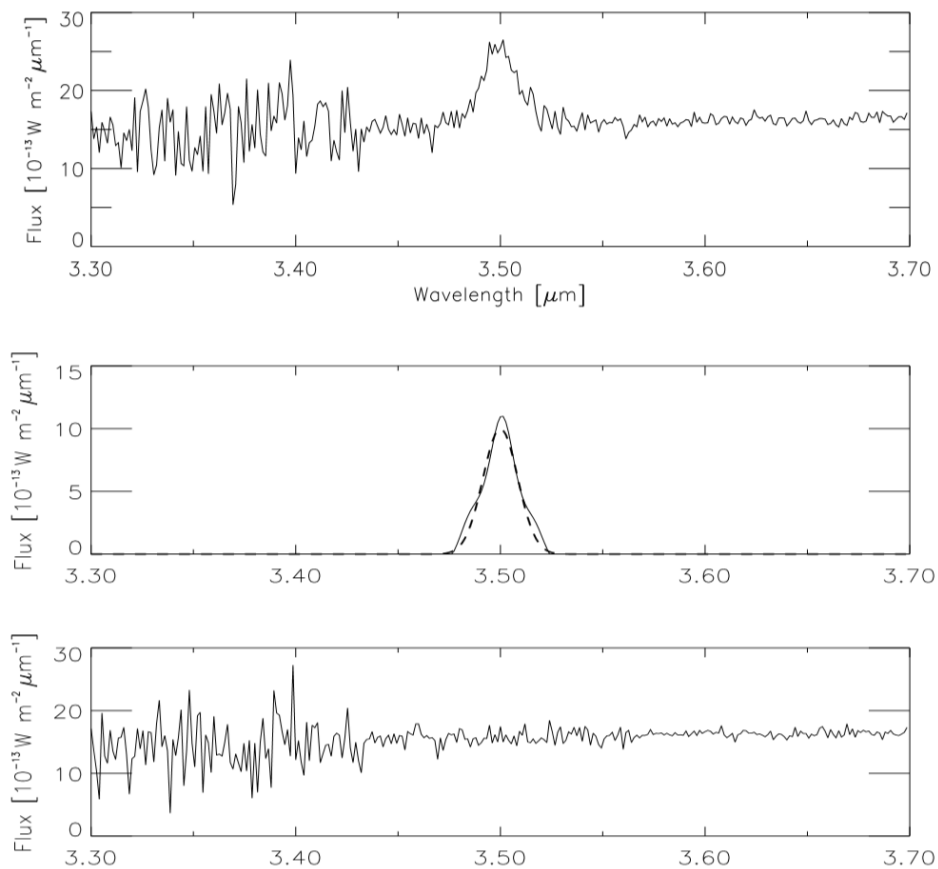


Рис. 9: Сверху моделированный спектр. В середине восстановленный спектр(сплошная линия), оригинальный моделированный спектр(пунктирная линия). Снизу остаточный спектр: разность первого и второго спектра. [1]

На рис. 10 в верхнем левом углу изображена галактика, которую необходимо отчистить от звезд, которые наложились сверху на данный снимок. Результат обработки приведен в нижнем правом углу. На рисунке справа сверху и слева внизу

приведены объекты, которые были обнаружены методом на 1 и 2 уровне и с 3 по 6 соответственно, которые были удалены. На изображении (b) видно, что алгоритм удалил небольшую часть рукава галактики. Однако, это не сильно сказалось на общей картине и результат все равно впечатляет.

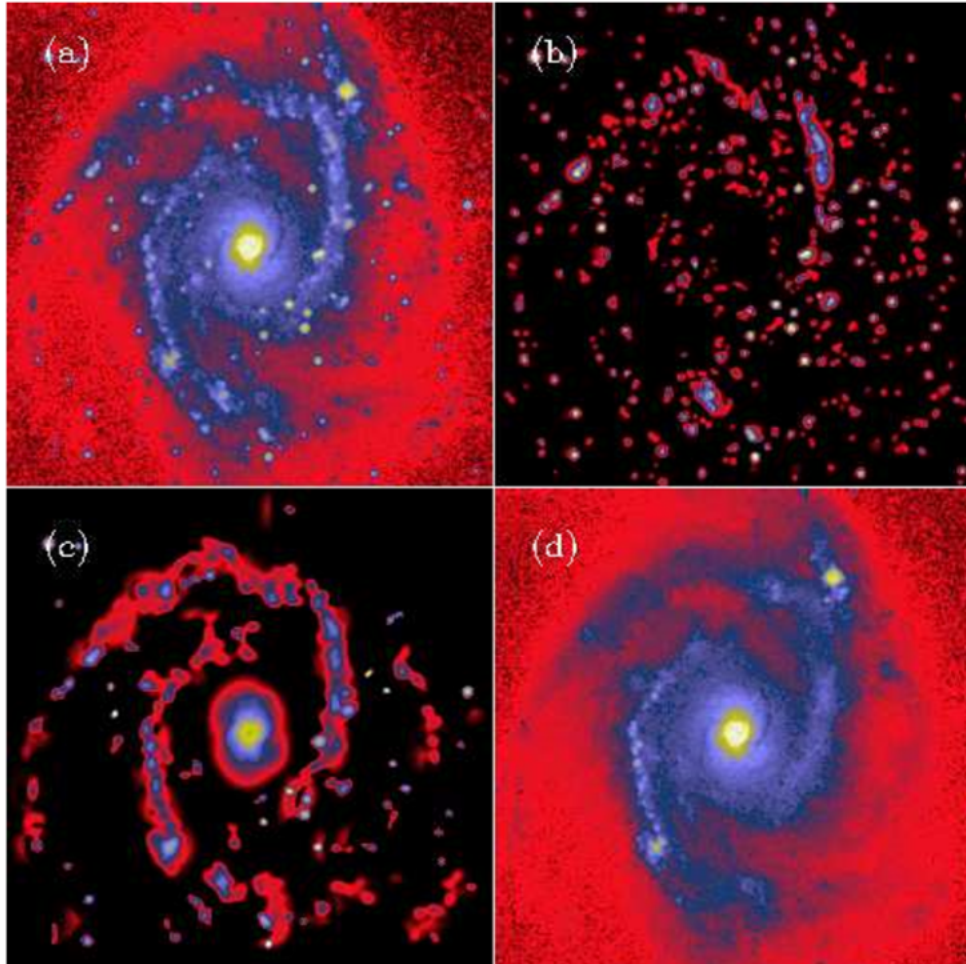


Рис. 10: (a) Галактика NGC2997, (b) найденные на уровнях 1 и 2 объекты, (c) найденные на уровнях 3,4,5,6 объекты, (d) разность между (a) и (b). [1]

На рис. 11 вверху слева приведено исходное изображение галактики. Можно заметить, что у ядра прослеживаются 2 коротких рукава, но более подробных морфологических наблюдений сделать не удастся. Но с помощью MVM подхода галактику можно разбить на её составляющие части и, например, выделить отдельно ядро и остальную часть галактики. После такого преобразования глазу становится заметна ранее скрытая за ярким ядром спиральная структура в центральной части галактики.

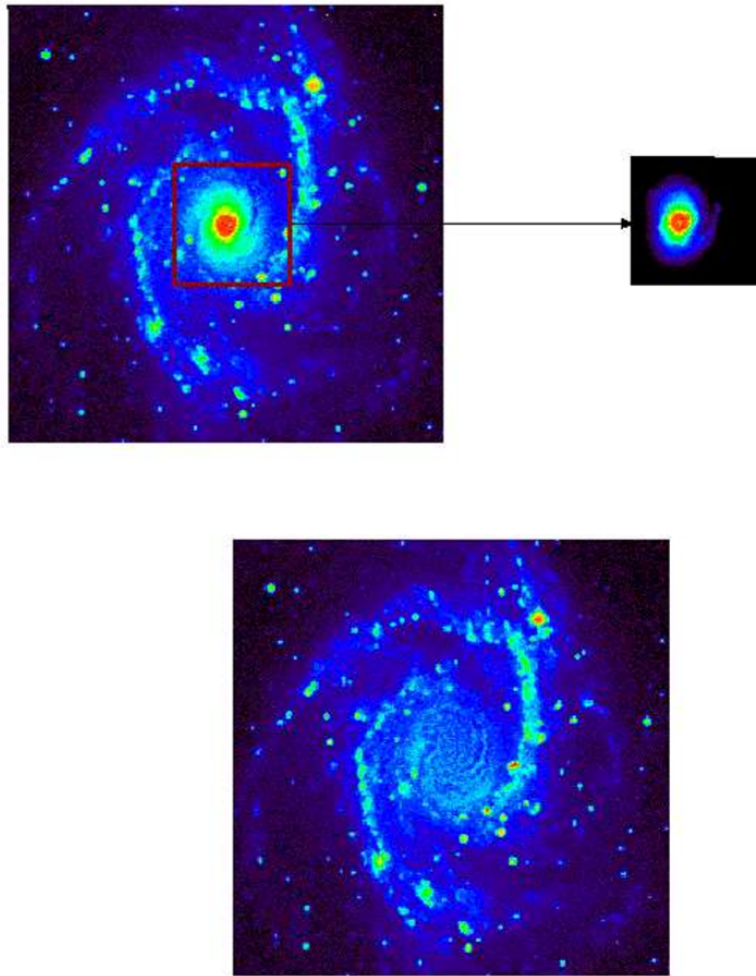


Рис. 11: В верхнем левом углу изображена галактика NGC2997, напротив нее, справа выделенное ядро, снизу разность между двумя предыдущими изображениями. [1]

6.9 Выводы

Описанный выше подход является инструментом для обширного исследования астро- снимков. Он преобразует большое изображение в сложное структурное объединение, которое позволяет как разделить несвязанные друг с другом объекты, так и компоненты внутри одного объекта. Подобная функциональность метода позволяет более детально исследовать сами объекты, как и произвести их классификацию или вычислить фотометрические показатели. При современном количестве поступающих с телескопов изображений, подобная автоматизация является ценной.

7 Конкурс Galaxy Zoo

В данном разделе будут описаны собственные эксперименты автора, направленные на решение задачи на платформе kaggle конкурса Galaxy Zoo [9]. Стояла задача морфологической классификации галактик (точнее, поиск вероятности принадлежности галактики к каждому из классов) по их изображениям. Более подробно о задаче можно почитать в [9].

Сразу хочется отметить, что из-за малого количества времени, которое оставалось до конца конкурса, не все гипотезы и идеи удалось проверить на практике. Для части результатов не осталось хороших иллюстраций и примеров. Поэтому далее будет описан один из возможных подходов к решению данной задачи, который получился в итоге множества экспериментов. Будет приведено сравнение некоторых подходов, где оно было реализовано, однако, исключительно как дополнительная часть, а не основная.

7.1 Специфика данных

Данные, которые предоставили организаторы конкурса были достаточно низкого качества. Изображения размером 424 x 424 пикселей, зашумленные не только обычным шумом, но и другими космическими объектами, объект интереса находится в центре, занимает на нем около 10%, его границы размыты, а ориентация случайная. По некоторым изображениям проходят черные полосы, которые, видимо, являются артефактами прибора, на который производилась съемка галактик. Все это многократно затрудняет работу с данными. Алгоритмы, которые заточены на точные границы и четкие изображения в данном случае не сработают. Приходится искать другие подходы, учитывающие вышеописанные трудности.

7.2 Схема решения

Изначально был выбран определенный принцип решения, который до конца конкурса остался неизменным: извлечь из изображения информативные признаки для каждого класса и применить для получившегося признакового пространства один из

наиболее универсальных и эффективных методов в наши дни, а именно Random forest.

Хочется отметить, что это оказался не самый лучший принцип, с точки зрения эффективности решения задачи. В некоторый момент метод уперся в свой потолок, который было тяжело преодолеть. Многие участники использовали нейронные сети для решения этой задачи, и в частности сверточными нейронными сетями, которые показали самый лучший результат в соревновании. Алгоритм, занявший 2 место, описан в [10].

Мой подход к задаче был разделен на 3 этапа, далее каждый этап будет рассмотрен подробнее.

1. **Предобработка изображений** В первую очередь имеется ввиду выделение объекта интереса и приведение его к некоторой нормальной форме.
2. **Извлечение признаков** После первого этапа можно выделять менее зашумленные и качественные признаки различной природы из изображений.
3. **Настройка метода** Настройка параметров Random forest и итоговая классификация тестовых объектов.

7.2.1 Предобработка изображений

Как уже было сказано основной целью данного этапа было желание привести все изображения к единому нормальному виду. Под нормальным в данном случае подразумевается следующий вид:

1. Большую часть изображения должен занимать объект интереса(галактика)
2. Потерять как можно меньше информации о галактике
3. Минимальное количество посторонних космических объектов на изображении и прочего шума
4. Ориентация вытянутой части галактики вдоль горизонтали

Основной задачей оказалось выделение галактики на изображении как вероятностной маски. Значение пикселя маски показывало вероятность того, что пиксель

исходного изображения принадлежит галактике. Дальнейшая работа не составляла труда и будет описана ниже. Было реализовано 2 подхода:

Бинаризация изображения по порогу. Порог искался по черно-белому изображению с помощью метода Отса и умножался на 1.3(экспериментально подобранная константа). Затем выбиралась центральная связанная область за бинарную маску галактики.

Метод оказался хорошим благодаря своей простоте, понятности и скорости работы. По началу он давал приемлемый результат, но обладал существенными недостатками. Он мог удалить информативную часть галактики на ее границе. Если по галактике проходила черная полоса, в конечном итоге оставалась одна из частей галактики. В редких случаях центральный пиксель не принадлежал объекту интереса и выделялся не тот объект.

Этот метод был различными способами модифицирован, что улучшило его качество. Например, поиск маски велся по сильно сглаженному изображению медианным фильтром, который убирал черные полосы на изображении(при этом дальнейшая работа велась с нефильтрованным изображением, чтобы не терять информативность). Или пиксели, которые находились недалеко от полученной маски, но не входили в нее получали значение между 0 и 1, которое уменьшалось с удалением от единиц маски. Однако, желаемый результат не был достигнут.

Поэтому был придуман другой метод, который был назван градиентным. В нем использовалась особенность данных изображений. Центр галактики всегда самый яркий и по удалению от него яркость как правило убывает. Соответственно, если найти центр галактики, то от него поиском в ширину пикселей с меньшим значением яркости относительно текущего пикселя можно выделить всю галактику. Если яркость опускалась ниже некоторого порога(это означало явную непринадлежность точки к галактике), то дальнейший поиск от данного пикселя прекращался. Это основная идея метода.

Поиск маски снова шел по сильно сглаженному изображению медианным и гауссовым фильтрами и разрешался рост по яркости, но не более чем на 0.5-1.5%. Затем найденная область сглаживалась таким образом, что пиксели около границы получали промежуточное значение между 0 и 1.

При таком подходе помимо основных требований достигался еще один интересный эффект. Если на изображение галактики накладывался некий выделяющийся яркий объект, например, звезда, то она затемнялась. Изображение галактики становилось более “чистым”. Пример работы метода показан на рис. 12.

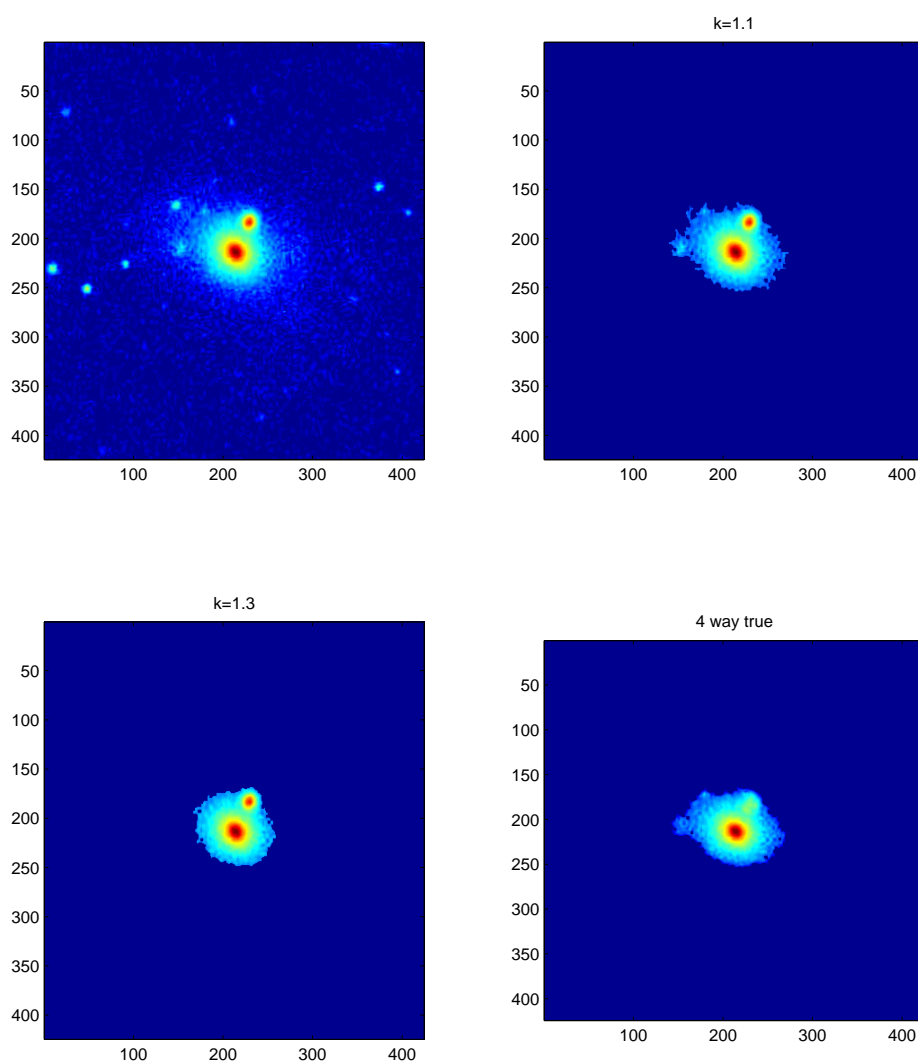


Рис. 12: Пример работы различных методов выделения галактики. Вверху слева — исходное изображение. Вверху справа и внизу слева — поиск с помощью отсеечения по порогу с разными константами. Внизу справа — градиентный способ выделения. Видно, что объект рядом с галактикой оказался затемненным, т.к. к самой галактике не относится.

Отметим, что качество работы данных алгоритмов оценивалось субъективно и визуально, поскольку никакой информации о том, как именно должна выглядеть хорошо выделенная и преобразованная галактика не было, поскольку это промежуточный этап задачи.

Полностью работу алгоритма описывает следующая схема:

1. Создание вероятностной маски одним из описанных методов
2. Попиксельное умножение исходного изображения на маску.
3. Поворот изображения. Выполнялся с помощью PCA. Координаты ненулевых пикселей рассматриваются как двумерное признаковое пространство. С помощью взвешенного PCA вычисляется угол поворота изображения. В качестве веса берется яркость пикселя.
4. Обрезка изображения по максимальным и минимальным координатам ненулевых пикселей.
5. Масштабирование изображения к выбранному размеру (например, 100 на 100 пикселей) с сохранением пропорций.

7.2.2 Извлечение признаков

Если метод предобработки изменился всего один раз, то набор разнообразных выдуманных признаков постоянно пополнялся и изменялся. В итоге было построено около 150-250 признаков. Ниже будут приведены некоторые принципиально разные классы признаков, которые использовались.

1. Цветовые признаки.

Не смотря на то, что классификация основана на сугубо морфологических особенностях объекта, цвет являлся достаточно важным признаком. Вычислялся средний цвет по центральной части изображения (3 признака) и еще логарифм от отношения зеленого к голубому и голубого к красному (2 признака). Отмечу, что часто помимо самих признаков входил признак, который являлся отношением двух исходных признаков, где это казалось уместным.

2. Признаки формы, яркости и пропорций галактики

Изображение делилось на несколько градаций по яркости: яркость больше 0.1, больше 0.3 и т.д. У подобных фрагментов галактики считались площадь, длина, ширина, дисперсия яркости

3. Признаки линий уровня галактики

Один из наиболее интересных признаков. Изображение галактики вытягивалось до совпадения высоты с шириной, потом переводилось в полярную систему координат с центром в центре галактики и вычислялись различные линии уровня по яркости либо по определенной яркости, либо по определенному расстоянию от центра галактики. Получалась некоторая кривая, которая сглаживалась и служила основой для очередной группы признаков. Вычислялись количества минимумов и максимумов кривой, ее длина, ее разброс значений.

После чего для галактики с помощью одной из линий уровня проводилось нелинейное преобразование в полярных координатах, которое выравнивало эту линию уровня. т.е. галактика “округлялась” таким образом, чтобы на определенном расстоянии была одна и та же яркость. После этого подобным образом обрабатывались линии уровня преобразованной галактики. Пример преобразований можно увидеть на рис. 13.

Таким образом была предпринята попытка выделения признаков для поиска спиральных структур. Некоторые галактики, которые были повернуты к Земле боком, после того, как происходило растяжение, выглядели ромбовидно, для того, чтобы поправить это искажение применялось нелинейное преобразование.

4. Признаки симметрии галактики

Признаки основанные на попарном сравнении 4х частей изображения, на которые делят проходящие через центр вертикальная и горизонтальная прямые. Считается корреляция и MSE.

5. PCA признаки

Первые коэффициенты в разложении изображения галактики по базису, который подбирает PCA. Вычислялись для непреобразованных изображений, для

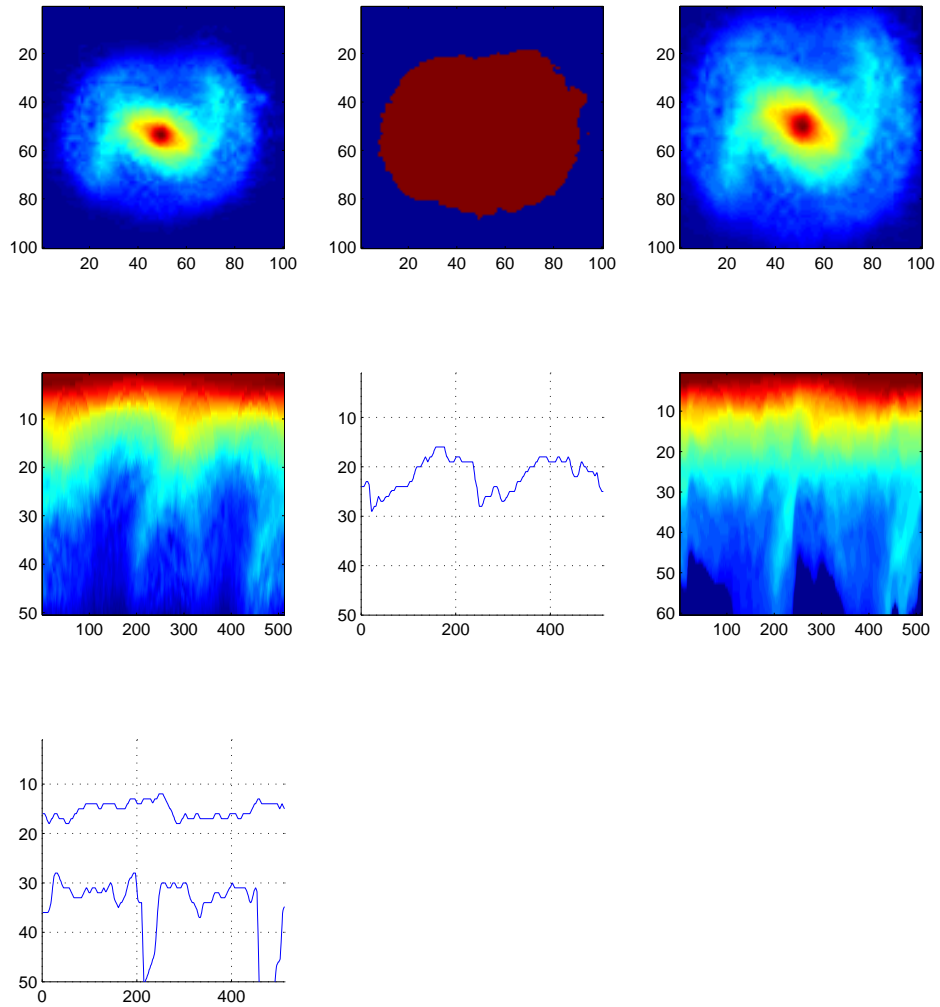


Рис. 13: Выделение признаков уровня галактики. Слева направо, сверху вниз: исходное предобработанное изображение, выделение галактики, растянутая галактика, так чтобы её длина и ширина оказались равны, предыдущее изображение в полярных координатах, линия уровня, по которой будет вестись выравнивание, выравненная галактика, линии уровня выравненной галактики

изображений в полярных координатах и нелинейно преобразованных, описанным выше способом.

7.2.3 Настройка метода

Была использована реализация Random forest, взятая из Matlab и настраивались два параметра: количество признаков, из которых выбирается следующее разбиение в текущем дереве и доля данных, которая используется для построения дерева. Эти параметры могут быть подобраны стандартными методами. Количество деревьев не является настраиваемым параметром, т.к. чем их больше, тем лучше.

По началу такой настройкой и ограничивалась вся работа. Однако, когда признаков стало много и встала необходимость тонкой настройки метода, была реализована следующая идея.

По ходу экспериментов было замечено, что если признаков слишком много, качество классификации падает. Поэтому было решено отобрать некоторое число максимально информативных признаков. Информативность также определялась Random forest'ом. Хотя это и не совсем правильный шаг, в данной задаче это не играло большой роли и к переобучению не приводило. Напомню, что в задании требовалось найти вероятности принадлежности к каждому из классов. Всего их 37. По большому счету, решалась задача регрессии и для каждого из ответов лучше всего подходили свои признаки. Поэтому для каждого ответа были подобраны свои лучшие признаки и их оптимальное количество.

Такая схема сработала и позволила повысить качество классификации.

8 Благодарности

Выражаю благодарность Воронцову Константину Вячеславовичу за прекрасное справочное пособие по latex, которое помогло оформить курсовую работу. А также Александру Геннадьевичу Дьяконову за хорошие и грамотные советы, которые очень помогли при решении задачи и написании курсовой работы.

Список литературы

- [1] J.L. Starck, F. Murtagh Handbook of Astronomical Data — 2002 Analysis
- [2] Kyle W. Willett, Chris J. Lintott Galaxy Zoo 2: detailed morphological classi

cations for 304,122 galaxies from the Sloan Digital Sky Survey

- [3] E. Bertin, S. Arnouts SExtractor: Software for source extraction — Astronomy & Astrophysics Supplement, 1996
- [4] Bertin, E. and Arnouts, S.: 1996, Astronomy and Astrophysics, Supplement Series 117, 393
- [5] Bijaoui, A.: 1980, Astronomy and Astrophysics 84, 81
- [6] Kron, R. G.: 1980, Astrophysical Journal Supplement Series 43, 305
- [7] Takase, B., Kodaira, K., and Okamura, S.: 1984, An Atlas of Selected Galaxies, University of Tokyo Press
- [8] Bijaoui, A. and Ru?e, F.: 1995, Signal Processing 46, 229
- [9] <http://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge>
- [10] https://github.com/milakov/nnForge/blob/master/examples/galaxy_zoo/galaxy_zoo.pdf?raw=true