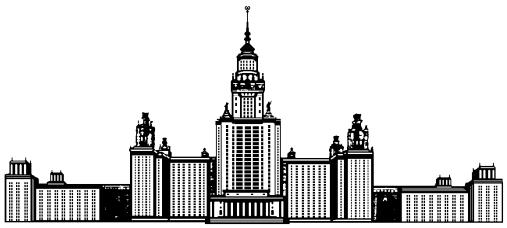


Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Даулбаев Талгат Кайратулы

Исследование принципов кластеризации карты звёздного неба

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:
чл.-корр. РАН
К.В. Рудаков

Москва, 2016

Содержание

1 Введение **3**

2 Постановка задачи **3**

3 Решение задачи **5**

3.1 Изучение данных 5

3.1.1 О равномерности выбора точек на сфере 5

3.1.2 Угол между тройками звёзд в одном созвездии 6

3.2 Самая яркая звезда в созвездии 7

3.2.1 Модули разностей видимых блесков 8

3.2.2 Среднее расстояние между звёздами 9

3.2.3 Выводы 9

3.3 Классификация рёбер 10

3.3.1 Логистическая регрессия 10

3.3.2 Случайный лес 11

3.3.3 Градиентный бустинг над решающими деревьями 12

3.3.4 Выводы 13

4 Трёхмерная визуализация **14**

5 Заключение **15**

1 Введение

В современной астрономии под созвездиями понимаются участки, на которые разделена небесная сфера для удобства ориентирования на звёздном небе. Некоторые из них насчитывают до нескольких сотен видимых звёзд.

Однако в древности во взаимном расположении звёзд наблюдатели видели некую систему и объединяли некоторые звёзды в одну группу, называя их созвездиями. Более того, в понимании древних людей созвездия были не просто множеством звёзд, а множеством звёзд и рёбер, которыми соединялись пары звёзд. Таким образом, они имели графовую структуру. Однако компонент связности было меньше, чем созвездий, потому что некоторые звёзды могли принадлежать сразу нескольким созвездиям. Так, например, альфа Пегаса была ещё и частью Андромеды.

В начале XIX века на карте звёздного неба созвездиями начали считать уже не большую группу звёзд, а целые участки звёздного неба так, что любая точка небесной сферы стала принадлежать какому-либо созвездию. Однако полной договорённости о том, как именно проходят границы между созвездиями, достичь так и не удалось.

Лишь в 1922 году на I Генеральной ассамблее Международного союза был окончательно принят список из 88 созвездий, на которые разбивается звёздное небо, а в 1928 году между ними были утверждены чёткие границы, проведённые строго по дугам небесной сферы. Наконец, в 1935 году после нескольких лет уточнений астрономы договорились, что больше никаких изменений вносить они не будут.

Изначально кажется, что выбор того, какие звёзды объединять в созвездия, был исключительно субъективным. Возможно, по этой причине работ по исследованию принципов кластеризации звёздного неба нет. Однако существуют некоторые работы (например, [3]), в которых задаётся вопрос, что было бы, если звёзды были кластеризованы современными алгоритмами машинного обучения. При этом авторы не стремятся понять принципы выделения созвездия, а лишь сравнивают результат кластеризации с настоящими созвездиями, получая низкое значение метрики Adjusted Rand Score.

2 Постановка задачи

Рассмотрим выборку из 681 звёзды, каждая из которых принадлежит одному из 88 созвездий. Здесь и далее под созвездием мы будем понимать характерную фигуру

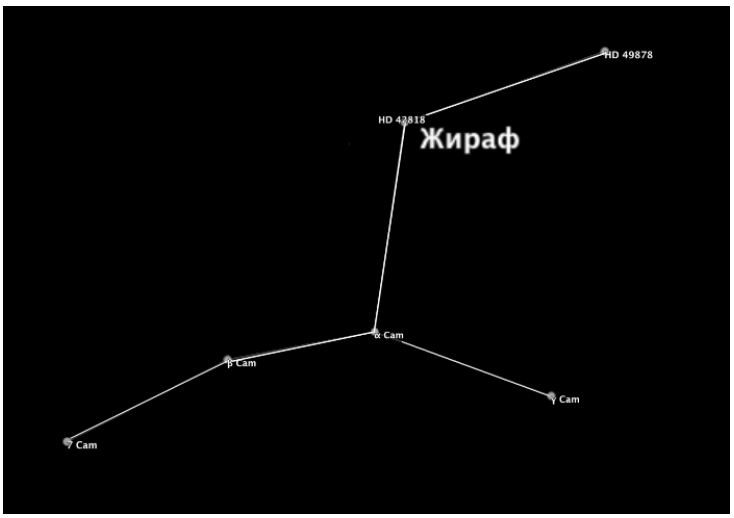


Рис. 1: Созвездие Жираф

на звёздном небе, как это делали люди древности. Множество этих звёзд обозначим за $\mathcal{V}$.

Множество всех пар звёзд одного созвездия обозначим за $\widehat{\mathcal{E}}$.

Как уже было сказано, некоторые пары звёзд древние люди соединяли линией на небесной сфере (например, см. рис. 1). В астрономии нет определения таких звёзд, поэтому будем называть две звезды, между которыми проведено ребро, непосредственно связанными. Множество пар непоследственно связанных звёзд обозначим за $\mathcal{E}$. Данные о них были найдены в книге [2].

Кроме того, рассмотрим выборку из всех звёзд, видимых невооружённым глазом. Этих звёзд оказалось 8707.

Для каждой звезды известны:

- Прямое восхождение α, или первая экваториальная координата — размер дуги небесного экватора от точки весеннего равноденствия до круга склонения светила.

- Склонение δ, или вторая экваториальная координата — угловое расстоянию на небесной сфере от плоскости небесного экватора до светила.

- Видимая звёздная величина m — мера яркости небесного тела (точнее, освещённости, создаваемой этим телом) с точки зрения земного наблюдателя.

  В выборке эта величина принимает значения от -1.0 до 5.47. Звёзды с $m > 6.5$ не видны невооружённым глазом.

Далее видимую звёздную величину будем называть видимым блеском звезды, как это принято в астрономии.

- Созвездие, которому принадлежит звезда.
- Список звёзд, с которыми звезда непосредственно связана в созвездии.

Целью работы является исследование того, по каким принципам звёзды выделяли в созвездия и почему человек соединял некоторые пары звёзд друг с другом.

Эта задача не из области физики, а скорее из области психофизиологии человека.

3 Решение задачи

3.1 Изучение данных

3.1.1 О равномерности выбора точек на сфере

Таблица 1: Количество квадратов $5° \times 5°$ с определённым количеством звёзд

Число звёзд в квадрате	0	1	2	3	4	5	6	7	8	9...
Реальные данные	685	650	494	293	222	122	48	45	14	19
Пуассоновское приближение	384	733	700	445	212	81	25	7	1	0

Самая тусклая звезда среди тех, что древние люди относили в созвездия, имеет блеск, приблизительно равный 6,0.

Не ограничиваясь созвездиями, рассмотрим абсолютно все 4899 звёзд, видимая звёздная величина которых не превышает этой величины и исследуем, насколько равномерно яркие звёзды распределены на небесной сфере. Для этого разобьём небесную сферу на квадраты $5° \times 5°$ и вычислим, сколько звёзд лежит в каждом из них (таблица 1).

Таких квадратов всего $72 \times 36 = 2592$, а звёзд всего 4945. Если бы звёзды с блеском не меньше 6.0 были распределены случайно, то распределение числа квадратов с заданным количеством звёзд можно было бы приблизить распределением Пуассона с параметром $\lambda = \dfrac{4945}{2592}$, но, очевидно это не так (см. таблицу 1).

Значит, мир устроен так, что звёзды на небесной сфере не распределены равномерно.

3.1.2 Угол между тройками звёзд в одном созвездии

Будем рассматривать тройки звёзд в одном созвездии такие, что одна звезда из тройки непоследственно связана с оставшимися двумя. Другими словами, рассмотрим все такие пары $((v_1, v_3), v_3)$, что $(v_1, v_2) \in \mathcal{E}$, и $(v_2, v_3) \in \mathcal{E}$.

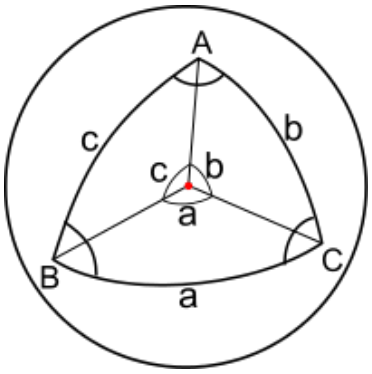


Рис. 2: Сферический треугольник

Для каждой такой тройки звёзд вычислим образованный ими угол A. Сделать это можно с помощью теоремы косинусов на сфере, из которой следует, что

$$A = \arccos \frac{\cos a \, - \, \cos b \, \cos c}{\sin b \, \sin c},$$

где a, b, c — это расстояния между звёздами в градусах.

В свою очередь расстояние в градусах между двумя звёздами с координатами (α_1, δ_1) и (α_2, δ_2) может быть вычислено по формуле:

$$d = 2 \arcsin \sqrt{\sin^2 \left(\frac{\Delta\delta}{2} \right) + \cos \delta_1 \cdot \cos \delta_2 \cdot \sin^2 \left(\frac{\Delta\alpha}{2} \right)},$$

где $\Delta\delta = \delta_1 - \delta_2$, $\Delta\alpha = \alpha_1 - \alpha_2$.

Вычислим углы для всех заданных троек звёзд и построим их распределение, сгладив гистограмму гауссовым ядром (3). Вычислив площадь под графиком, получим, что всего 22% углов острые.

Усреднив полученные значения углов по созвездиям, получим распределение с центром в точке 120° (рис. 4).

Теперь рассмотрим все тройки звёзд и усредним углы между ними, тогда получим значение около 89°.

Этот факт даёт сделать предположение, что при выделении экстраполяции связей между звёздами человек избегает острых углов.

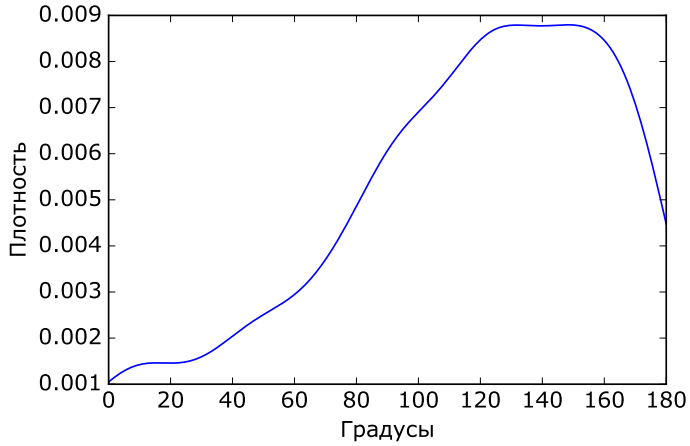


Рис. 3: Распределение градусов углов между рёбрами.

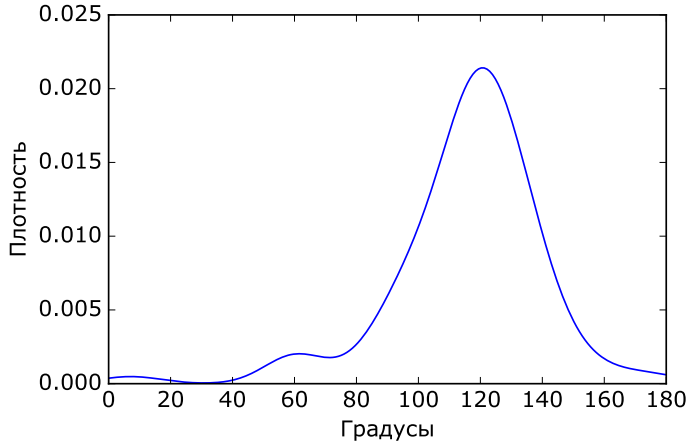


Рис. 4: Распределение градусов углов между рёбрами.

3.2 Самая яркая звезда в созвездии

Можно предположить, что созвездия всегда образуются вокруг одной яркой звезды, чтобы проверить это построим гистограмму (рис. 5).

Среднее значение видимого блеска звезды по всем звёздам, принадлежащим всем созвездиям, равно 3.58, что соотносится со средним видимым блеском самой яркой звезды в созвездии.

Поэтому предположение о том, что в каждом созвездии должна быть яркая звезда, несправедливо.

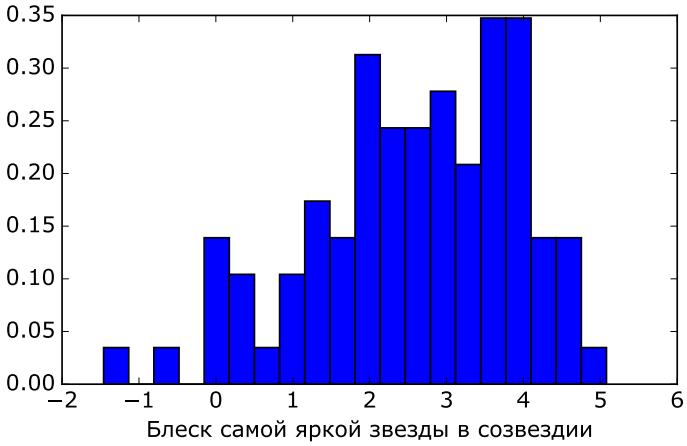


Рис. 5: Гистограмма видимого блеска самой яркой звезды в созвездии.

3.2.1 Модули разностей видимых блесков

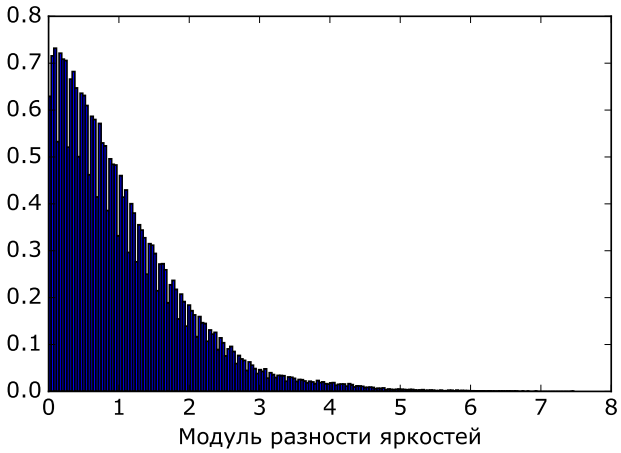


Рис. 6: Распределения модулей разностей яркостей между всеми парами звёзд.

Изучим распределения модулей разности видимых блесков звёзд (рис. 6). Можно заметить, что оно очень похоже на распределение модуля нормальной случайной величины.

Однако если наблюдения отразить симметрично относительно оси ординат, то критерий Шапиро — Уилка отвергнет нормальность с p-value равным нулю.

Рассмотрим теперь среднее значение модуля разности видимых блеском звёзд, если считать созвездие полным графом (см. рис. 7).

Распределения заметно отличаются от того, что представлено на рис. 6.

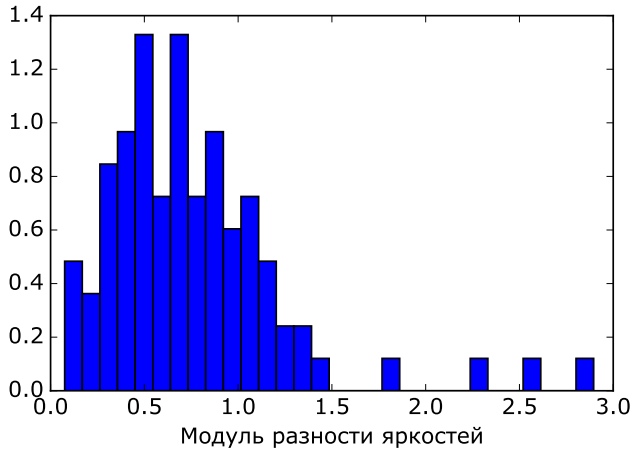
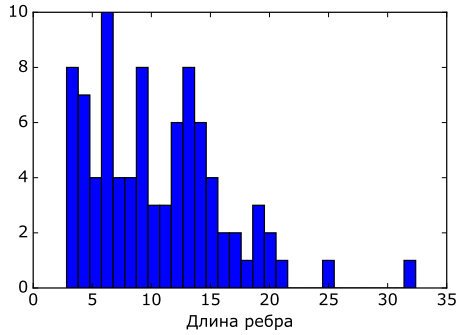


Рис. 7: Распределения модулей разностей яркостей между непосредственно связанными звёздами

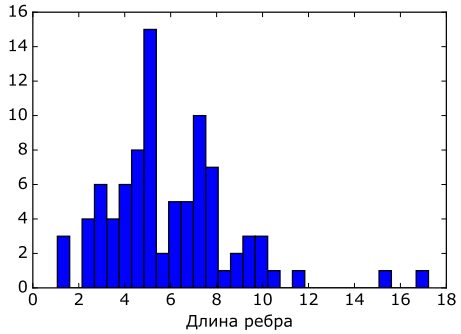
3.2.2 Среднее расстояние между звёздами

Среднее расстояние между всеми парами звёзд оказалось равным 88.9°, что соответсвует нашим представлениям. В свою очередь, среднее расстояние между непосредственно связанными звёздами равно 5.97°, а среднее расстояние между двумя звёздами одного созвездия равно 10.51°.

Гистограммы распределений приведены на рисунке 3.2.2.



Распределение длин рёбер, если созвездия имеют структуру полного графа.



Распределение длин рёбер между всеми непосредственно связанными звёздами.

3.2.3 Выводы

Несмотря на то, что выделенные созвездия кажутся неструктурированными и хаотичными, существуют признаки, которые отличают звёзды одного созвездия — в частности, непосредственно связанные звёзды — от пар звёзд разных созвездий.

3.3 Классификация рёбер

Сведём задачу к классификации пар рёбер. Ответом будет являться то, является ли пара рёбер непосредственно связанной.

В каждой паре первой будем называть звезду с наибольшей яркостью в паре, если яркости звёзд совпадают, то будем называть первой произвольную.

Для каждой пары (u_i, u_j) звёзд введём следующие признаки:

dist: угловое расстояние между двумя звёздами, измеряющееся в градусах и принимающее значение от $0°$ до $180°$;

m_i: яркость первой звезды, увеличенная на единицу;

m_j: яркость второй звезды, увеличенная на единицу;

h_i: бинарный признак, отвечающий за то, видна ли первая звезда из северного полушария;

h_j: бинарный признак, отвечающий за то, видна ли вторая звезда из северного полушария;

n_i: количество звёзд в окрестности $5°$ первой звезды;

n_j количество звёзд в окрестности $5°$ второй звезды;

5NNd_i: угловое расстояние до пяти ближайших соседей первой звезды (вектор длины 5);

5NNd_j: угловое расстояние до пяти ближайших соседей второй звезды (вектор длины 5);

Далее попробуем применить несколько стандартных алгоритмов машинного обучения.

3.3.1 Логистическая регрессия

Отмасштабируем данные, а затем обучим на них логистическую регрессию — линейный классификатор, способный выдавать оценку вероятности принадлежности объекта каждому классу. Логистическая регрессия также позволяет оценить важность признаков: если предварительно масштабировать признаки, абсолютное значение коэффициента при каждом признаке будет говорить о том, насколько он значим.

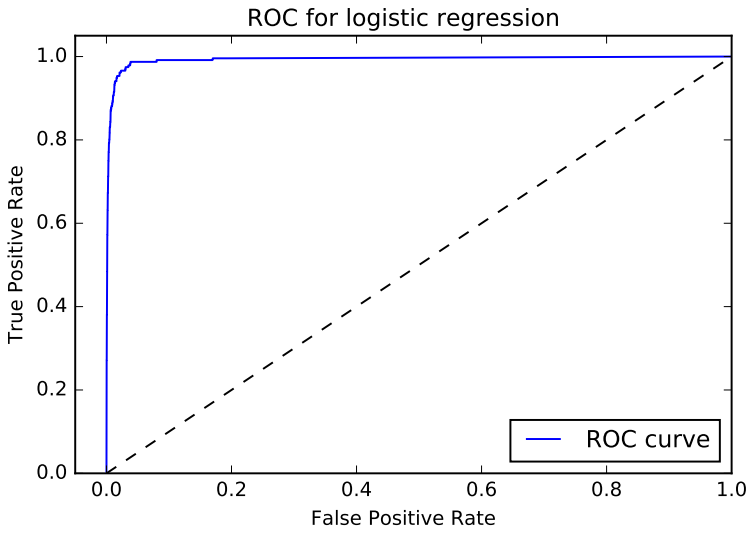


Рис. 8: ROC-кривая на тестовой выборке для логистической регрессии.

На рис. 8 представлена ROC-кривая для логистической регрессии, полученное значение AUC — 0,993.

Полученные коэффициенты:

$$\underbrace{-18.582}_{\text{dist}}\ \underbrace{-0.377}_{m_i}\ \underbrace{0.314}_{m_j}\ \underbrace{0.065}_{h_i}\ \underbrace{0.0}_{h_j}\ \underbrace{-0.012}_{n_i}\ \underbrace{0.057}_{n_j}$$

$$\underbrace{0.219, 0.117, 0.078, 0.063, 0.209}_{5\text{NNd}_i}\ \underbrace{0.26, 0.107, 0.08, 0.198, -0.009}_{5\text{NNd}_j}$$

Как можно было ожидать, самый важный признак — расстояние между звёздами. Наименее важные признаки — полушарие, количество ближайших соседей в окрестности.

3.3.2 Случайный лес

Проведём такие же исследования с алгоритмом случайного леса. Этот алгоритм является композицией решающих деревьев, обучаемых независимо. Каждое решающее дерево обучается по своей подвыборке, полученной из исходной с помощью «выбора с возвращением», таким образом в обучающей выборке будут встречаться одинаковые объекты. Кроме того, при построении каждой новой внутренней вершины деревьев выбирается случайное подмножество признаков так, что оптимальное значение признака выбирается из подмножества, а не из всего множества.

Этот классификатор тоже способен строить оценки важности признаков. Базируется этот алгоритм на простой идее, что признак важен, если при случайной перестановке значения этого признака в выборке, ошибка сильно возрастёт. Под ошибкой

понимается так называемая out-of-bag ошибка, которая вычисляется для каждого дерева по тем объектам, которые не попали в выборку.

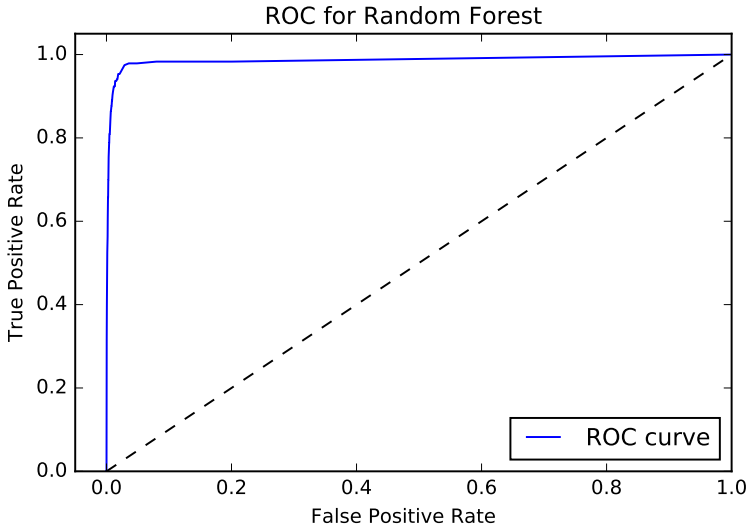


Рис. 9: ROC-кривая на тестовой выборке для случайного леса.

На рис. 9 представлена ROC-кривая для случайного леса, полученное значение AUC — 0.987.

Оценки важности признаков:

$$\underbrace{0.543}_{\text{dist}}\ \underbrace{0.034}_{m_i}\ \underbrace{0.038}_{m_j}\ \underbrace{0.005}_{h_i}\ \underbrace{0.004}_{h_j}\ \underbrace{0.009}_{n_i}\ \underbrace{0.009}_{n_j}$$

$$\underbrace{0.034, 0.038, 0.037, 0.037, 0.038}_{5\text{NNd}_i}\ \underbrace{0.034, 0.034, 0.035, 0.037, 0.034}_{5\text{NNd}_j}$$

Как мы видим, вновь длина ребра является самым важным признаком.

3.3.3 Градиентный бустинг над решающими деревьями

Другим популярным алгоритмом машинного обучения является градиентный бустинг над решающими деревьями. Этот алгоритм в отличие от случайного леса строит композицию решающих деревьев не независимо, а так, что каждый следующий базовый алгоритм учитывает ошибку предыдущих базовых алгоритмов.

На рис. 10 представлена ROC-кривая для случайного леса, значение AUC — 0,991.

Полученные оценки важности признаков:

$$\underbrace{0.241}_{\text{dist}}\ \underbrace{0.085}_{m_i}\ \underbrace{0.087}_{m_j}\ \underbrace{0.001}_{h_i}\ \underbrace{0.005}_{h_j}\ \underbrace{0.003}_{n_i}\ \underbrace{0}_{n_j}$$

$$\underbrace{0.064, 0.07, 0.062, 0.054, 0.049}_{5\text{NNd}_i}\ \underbrace{0.077, 0.048, 0.048, 0.044, 0.061}_{5\text{NNd}_j}$$

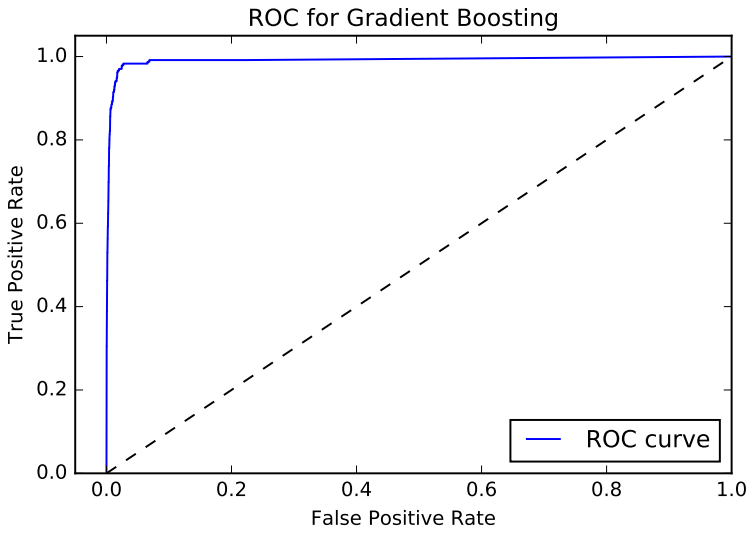


Рис. 10: ROC-кривая на тестовой выборке для случайного леса.

3.3.4 Выводы

Базовые алгоритмы классификации работают неплохо для выделения рёбер. Однако алгоритмы оценки важности признаков говорят, что расстояние между звёздами намного более важный признак, чем все остальные.

Таким образом, нам не удалось найти таких одиночных и парных признаков звёзд, кроме расстояния и яркостей, которые бы позволили отделить непосредственно связанные звёзды от несвязанных.

4 Трёхмерная визуализация

В рамках работы был реализован программный стенд для трёхмерной визуализации звёзд.

С его помощью можно как посмотреть на все созвездия, так и выделить узнать характеристики отдельной звезды.

Ниже представлены несколько скриншотов из программы.

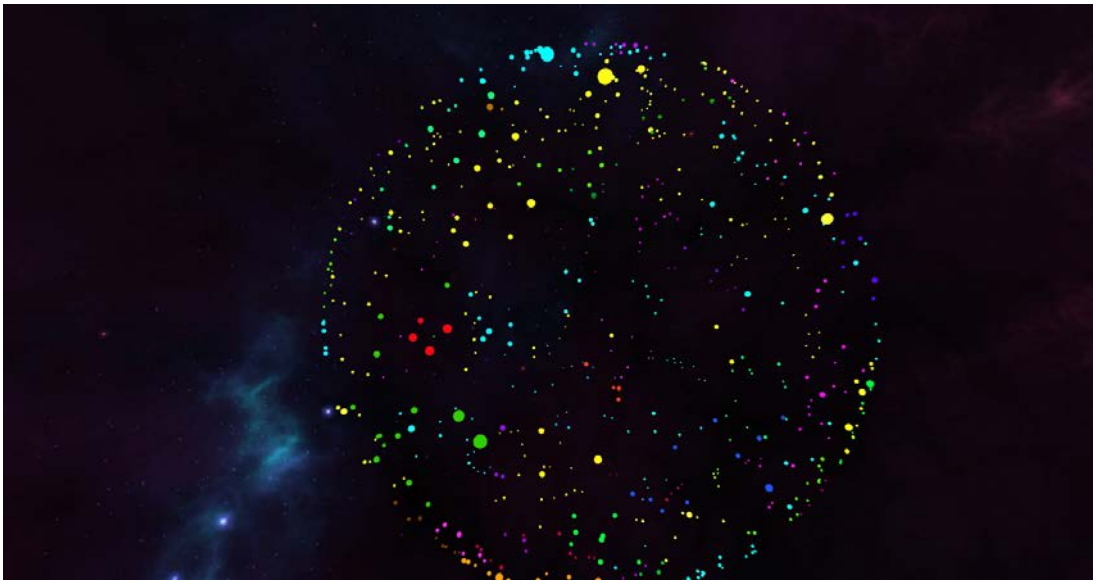


Рис. 11: Небесная сфера. Разные созвездия выделены разными цветами.

Рис. 12: Если выделить звезду, она окрасится красным цветом, а её созвездие — жёлтым, в левом углу появится информация.

5 Заключение

В результате проведённой работы:

1. Исследована предметная область;
2. Проанализированы некоторые признаки, характеризующие положение звёзд в созвездиях;
3. Установлено, что прямые попытки применить стандартные алгоритмы не приводят к должному результату;
4. Реализован программный стенд для визуализации звёзд;

Список литературы и ресурсов сети Интернет

[1] Yale Bright Star Catalog. — [Доступен 23-апрель-2016]. `http://tdc-www.harvard.edu/catalogs/bsc5.html`.

[2] *Barton, S. G.* A Guide To The Constellations / S. G. Barton, W. H. Barton. — 1928.

[3] *Xu, S.* Re-clustering of Constellations through Machine Learning / S. Xu, K. Chen, Y. Zhou. — 2014.

[4] *Кононович Э. В.,*. Общий курс астрономии / Кононович Э. В., Мороз В. И. — БИНОМ, 2011.

[5] *Лагутин М. Б.,*. Наглядная математическая статистика / Лагутин М. Б. — БИНОМ, 2013. — P. 472.