# Application of Machine Learning Methods for Cross-Identification of Astronomical Objects

Alexandra Kulishova

Lomonosov Moscow State University, Moscow, Russia
sasha_kulishova@mail.ru

**Abstract.** In the modern world the number of astronomical observations is steadily growing per day. It is necessary to identify observations from known astronomical sources to work effectively with this information. The article is devoted to the application of machine learning methods for cross-identification of astronomical objects. The paper presents a generalization of ideas from similar works. A brief comparison of the applied models is given. There are described setting and results of the experiment that was carried out on the basis of the analysis of other articles. This work can serve as a basis for the implementation of the stage of cross-identification of observations in astronomical systems.

**Keywords:** machine learning, cross-identification, astronomy

# Применение методов машинного обучения для кросс-идентификации астрономических объектов

Кулишова Александра

Московский государственный университет им. М. В. Ломоносова, Москва, Россия
sasha_kulishova@mail.ru

**Abstract.** В современном мире количество астрономических наблюдений в сутки неуклонно растет. Для эффективной работы с данной информацией необходимо идентифицировать наблюдения по известным астрономическим источникам. Статья посвящена применению методов машинного обучения для кросс-идентификации астрономических объектов. В работе представлено обобщение идей и результатов из схожих работ. Приводится краткое сравнение применяемых моделей. Описывается постановка и результаты эксперимента, проведенного на основе анализа других статей. Данная работа может послужить основой для реализации этапа кросс-идентификации наблюдений в астрономических системах.

**Keywords:** машинное обучение, кросс-идентификация, астрономия

# 1 Введение

Важной задачей в астрономии с давних пор является наблюдение за ночным небом, а точнее за излучением астрономических объектов. Именно благодаря наблюдениям выявляются новые источники излучения. Одной из задач в современной астрономии является классификация наблюдений на стационарные источники и транзиенты, а также проведение специальных длительных наблюдений отдельных транзиентов для уточнения их свойств. В результате наблюдений, проводимых в различных обсерваториях, получаются большие массивы астрономических изображений.

Астрономическое изображение, полученное с помощью современного телескопа, имеющего не очень большое поле зрения, может содержать несколько тысяч объектов. В наблюдениях, которые носят поисковый характер, необходимо получить астрометрическую и фотометрическую информацию для всех объектов на изображении. Обработать такое количество информации вручную невозможно, а рост получаемых данных в сутки неуклонно растет: например, современные телескопы позволяют собирать около 50 гигабайт (Gaia) данных о наблюдениях в сутки, а в скором времени и около 30 терабайт (LSST) в сутки.

Задача классификации источников и поиска транзиентов на основе анализа астрономических изображений может разбиваться на три подзадачи:

1. Каталогизация объектов на серии изображений одного участка неба в разные эпохи.
2. Кросс-идентификации полученных каталогов
3. Классификация источников и поиск транзиентов с помощью применения методов анализа кривых блеска.

Кросс-идентификация – это отождествление наблюдений с объектами, к которым они относятся. Эти объекты в зависимости от системы могут принадлежать, как к одному каталогу, так и к нескольким [5].

Этап кросс-идентификации необходим и имеет ряд сложностей, так как астрономические каталоги различаются по методам получения информации для них. Эти различия приводят к тому, что один и тот же физический источник в разных каталогах имеет отличающиеся характеристики. Например, наблюдения были зарегистрированы в разные эпохи и в разных спектральных диапазонах. Также инструменты наблюдения имеют разную чувствительность, которая дает различную астрометрическую и фотометрическую погрешность. Задача кросс-идентификации состоит в отождествлении наблюдений или объектов из разных каталогов, являющихся обнаружением одного и того же физического источника. Тем самым кросс-идентификация позволяет получить более полную информацию об источнике и отследить изменение характеристик с течением времени.

Одним из наиболее известных методов является Байесовский подход к кросс-идентификации астрономических объектов. Однако, с ростом популярности методов машинного обучения, появились работы, связывающие их с данной задачей. Это обусловлено тем, что алгоритмы машинного обучения могут работать

быстрее и с большим объемом данных. В данной работе рассматриваются способы применения методов машинного обучения для решения задачи кросс-идентификации в зависимости от постановки конкретной задачи.

## 2 Постановка задачи

Цель данной работы - применение методов машинного обучения для кросс-идентификации астрономических объектов. В рамках статьи были выполнены следующие подзадачи:

1 Анализ существующих подходов кросс-идентификации астрономических объектов на основе методов машинного обучения.
2 Выбор подходящих методов машинного обучения на основе результатов анализа.
3 Реализация выбранных методов: выбор данных, проведение экспериментов, выявление особенностей реализации.
4 Сравнительный анализ выбранных методов на наборах астрономических данных Gaia и PLAsTiCC.

### 2.1 Сценарии работы

В рамках работы [6] наблюдения получаются из изображений со съемок ночного неба. Для выделения объектов и их необходимых признаков набор изображений обрабатываются специальным образом. В статье был описан подробный метод каталогизации объектов. Результатом его применения является набор объектов с откалиброванными значениями признаков.

Однако в рамках поставленной задачи возможно несколько сценариев получения и обработки данных. Каждый из них зависит от источника входных данных, особенностей получения наблюдений и признаков каждого наблюдения. Все сценарии можно разделить на два типа: задачи классификации и кластеризации. Особенностью сценариев классификации является то, что нам заранее известен набор объектов (это может быть один или несколько каталогов), к которым мы хотим отождествить наблюдения. В рамках сценариев кластеризации каталог объектов может быть вообще не задан, а задачей будет являться группировка наблюдений для открытия новых небесных тел.

**Классификация.** В рамках задачи классификации необходимо сопоставить поступающие наблюдения с существующими объектами в заданном каталоге. Для этого строится модель классификатора, отождествляющая наблюдения с объектами из каталога на основе уже существующих объектов и наблюдений.

Поступающие в реальном времени данные предобрабатываются, а позже к ним применяется построенный классификатор, одной из особенностей которого является то, что он должен работать с любыми объектами, представленными в каталоге. Для этого надо либо обучить его на наблюдениях всех объектов (надо проверить осуществимо ли это по времени), либо разбить область неба на регионы и

построить классификаторы для каждого региона. Также можно использовать относительные координаты и обучить классификатор на представительной обучающей выборке, однако в этом случае возникает ряд проблем: высокая сложность подготовки обучающего набора, а также возможность того, что в этом случае методы классификации работать не будут из-за высокого количества классов.

**Кластеризация.** Необходимо сгруппировать поступающие наблюдения либо для их дальнейшего соотнесения с существующими объектами в каталоге, либо для добавления новых объектов в каталог. Для этого строим модель кластеризации наблюдений. Проверяем ее на известных наблюдениях, относящихся к известным объектам из каталога. Применяем алгоритм к новым наблюдениям, поступающим в реальном времени.

Особенностью кластеризации является то, что модель должна работать с наблюдениями любых объектов, представленных в каталоге. Для этого надо подобрать правильный алгоритм предобработки данных, а также параметры кластеризации.

## 3 Родственные работы

В рамках данной работы необходимо было выбрать методы машинного обучения, подходящие для кросс-идентификации наблюдений и астрономических объектов, а также выработать обобщенный алгоритм решения поставленной задачи. Для этого были проанализированы статьи, решающие схожую задачу.

### 3.1 L. Lindegren: Cross-Matching Gaia Objects

Так в своей статье [2] L. Lindegren, решая задачу перекрестного сопоставления для анализа данных в Gaia, предлагает свой алгоритм, основанный на выделении двух основных этапов:

1. Этап классификации: если задан набор исходных источников (например, каталог объектов), то применяется алгоритм классификации наблюдений к заданным объектам. Если остались наблюдения, которые не удалось классифицировать, то переходим на следующий этап.
2. Этап кластеризации: если набора исходных объектов нет (или наблюдения к ним не относятся), применяется алгоритм кластеризации.

В качестве метода классификации автор использует алгоритм К-ближайших соседей, а в качестве алгоритма кластеризации применяет иерархическую агломеративную кластеризацию. Однако в обычном алгоритме иерархической кластеризации возможно сопоставление всех пар объектов, что приводит к квадратичному увеличению времени вычислений при увеличении количества наблюдений. Поэтому в рамках оптимизации автором решено было использовать цепочки ближайших соседей (NNC) с заданным порогом, а также использовать в качестве функции расстояния расстояние Уорда, которое позволяет агломерировать

наблюдения с минимальной потерей информации. К этому же выводу пришли авторы статей [1] и [3], в которых подробно рассматривается задача кластеризации для сопоставления наблюдений и астрономических объектов.

Минусом данной работы является то, что тестирование проводилось только на симулированных данных. Для каждого наблюдения моделирование происходило в два этапа: сначала генерировалось случайное поле источников, а затем их наблюдения в определенные эпохи. Каждый источник был определен своим положением и величиной потока. В качестве оценки авторы использовали коэффициент успеха F, который равен 1, если все истинные пары были найдены, а ложных не было выявлено.

В первом тесте было сгенерировано случайное звездное поле с около 500 источниками без собственных движений. Позже плотность была изменена путем масштабирования размера поля. Позиционная неопределенность ($\sigma$) была зафиксирована 1 угловой секундой. Цель этих тестов - продемонстрировать возможность выполнить кластерный анализ данных на основе положений и правильных движениях. В результате автор показывает, что алгоритм сходится к правильному решению при уменьшении $\sigma$. Кроме того алгоритм также правильно распознал звезду с высоким собственным движением и дал значительно улучшенный показатель успеха (F = 0,898 против 0,810). При уменьшении $\sigma$ до 0,5 угловых секунд было получено правильное решение (F = 1). Конечно, такого мягкого поведения нельзя ожидать в сильно переполненных регионах.

Таким образом, тесты показали, что кластерный анализ может работать для сопоставления наблюдений и объектов, даже если движения сопоставимы с расстояниями между источниками.

### 3.2 D. J. Rohde, M. J. Drinkwater: Applying Machine Learning to Catalogue Matching in Astrophysics

К-ближайших соседей – не единственный применимый метод машинного обучения для такого рода задач. В своей работе [4] D. J. Rohde, M. J. Drinkwater подробно описывают метод выбора признаков и моделей машинного обучения для задачи сопоставления объектов из разных каталогов.

В рамках поставленной ими задачи необходимо было объединить каталог с относительно плохими позиционными неопределенностями (HIPASS) с каталогом с хорошими позиционными неопределенностями (SuperCOSMOS). Для этого был производен отбор необходимых признаков из заданных данных, а затем к ним применены такие алгоритмы классификации, как машина опорных векторов (SVM) и нейронные сети.

При отборе необходимых характеристик авторы предлагают искать корреляции между признаками объектов из разных каталогов, и проводить простую предобработку данных (например, удаление пустых данных). Чтобы выбрать наиболее подходящий метод машинного обучения и подбора его параметров, авторы использовали перекрестную проверку: данные делились на 10 равных частей, затем алгоритмы обучались на 9 из них и проверялись на 10-м. Эта процедура повторялась 10 раз с усреднением результатов.

Как видно из таблицы 1, наилучшим образом себя показал метод опорных векторов с полиномиальным ядром и степенью равной трем.

**Таблица. 1.** Результаты кросс-валидации рассматриваемых методов

| Algorithm (Kernel) | Soft margin (c)/hu | Positive data per cent correct | Negative data per cent correct | Overall Per cent Correct |
|---|---|---|---|---|
| SVM Linear | 0.1 | 87.47 ± 2.24 | 98.70 ± 0.27 | 97.17 |
| | 1 | 88.94 ± 2.43 | 98.75 ± 0.50 | 97.41 |
| | 10 | 88.80 ± 2.44 | 98.80 ± 0.25 | 97.43 |
| SVM Poly d = 2 | 0.1 | 90.04 ± 0.31 | 99.04 ± 1.67 | 97.93 |
| | 1 | 94.18 ± 1.91 | 99.44 ± 0.20 | 98.72 |
| | 10 | 96.02 ± 1.47 | 99.53 ± 0.29 | 99.05 |
| **SVM Poly d = 3** | 0.1 | 94.91 ± 1.93 | 99.46 ± 0.27 | 98.84 |
| | 1 | 96.24 ± 1.83 | 99.54 ± 0.20 | 99.09 |
| | **10** | **96.69 ± 1.26** | **99.50 ± 0.42** | **99.12** |
| SVM RBF γ = 1 | 0.1 | 89.39 ± 2.58 | 99.21 ± 0.27 | 97.87 |
| | 1 | 93.66 ± 2.47 | 99.50 ± 0.28 | 98.70 |
| | 10 | 95.43 ± 1.69 | 99.66 ± 0.17 | 99.08 |
| Perceptron | | 86.81 ± 7.73 | 97.52 ± 2.78 | 96.05 |
| Neural net | hu = 3 | 93.81 ± 1.68 | 95.50 ± 1.41 | 95.27 |
| | hu = 4 | 94.10 ± 2.07 | 95.46 ± 1.38 | 95.27 |
| | hu = 5 | 93.50 ± 3.59 | 95.48 ± 1.26 | 95.21 |
| | hu = 6 | 93.45 ± 2.01 | 95.62 ± 1.31 | 95.32 |

HOPCAT содержал 2221 объект, для которого было недостаточно информации для сопоставления. Именно из этих данных авторы извлекали новую информацию, сопоставив ее с помощью машинного обучения. В рамках данного решения удалось назначить уникальные совпадения для 1209 из них. Высокая точность тестового набора говорит о том, что очень высокая доля этих совпадений является правильной. Таким образом, D. J. Rohde и M. J. Drinkwater удалось увеличить количество совпадений двух заданных каталогов до 3096 из 4315 и показать, что для данного типа задач такие алгоритмы, как SVC и нейронная сеть, могут показать эффективный и качественный результат.

## 4 Анализ используемых методов

Для проверки эффективности применения методов машинного обучения для кросс-идентификации астрономических объектов на основе представленных ранее работ были выбрано несколько алгоритмов и проведен ряд экспериментов.

Для задачи классификации были выбраны алгоритмы, используемые в рассмотренных работах: К-ближайших соседей, метод опорных векторов (SVM), многослойный персептрон. Также рассматривалось решающее дерево (Decision Tree) для того, чтобы определить, как работает такая конструкция в представленной задаче. Реализация данного механизма проста, а скорость работы достаточно быстрая. С этой же точки зрения в качестве многослойного персептрона применялся MLPClassifier из библиотеки sklearn. При положительных результатах вместо них в будущей работе будут использоваться алгоритм случайного леса (Random Forest) и полноценно построенная нейронная сеть. В качестве алгоритмов кластеризации**:** агломеративная иерархическая кластеризация, К-средних (K-means), а также спектральная кластеризация.

В рамках данной работы были использованы реализации представленных алгоритмов из библиотеки sklearn [11].

### 4.1 Постановка задачи для экспериментов

Все рассмотренные выше сценарии в рамках представленных экспериментов имеют следующую постановку задачи: из каждого каталога PLAsTiCC и Gaia было выбрано одинаковое количество наблюдений, к которым независимо применялись алгоритмы классификации и кластеризации.

Все эти эксперименты делятся на несколько основных этапов:

1. Предобработка данных: на этом этапе происходит отбор необходимых для поставленной задачи признаков, обработка и модификация данных.
2. Разделения данных на обучающую и валидационную выборки. Данный этап необходим для качественной работы используемых алгоритмов, так как в выбранных источниках данных все наблюдения находятся в одной таблице и отсортированы в порядке их соотнесения к объекту.
3. Подбор параметров и обучение выбранного метода машинного обучения.
4. Получение показателей качества модели на валидационном наборе данных и сравнение с результатами работы других алгоритмов.

Данная постановка задачи имеет несколько обязательных условий:

1. При классификации классом является отдельный астрономический объект.
2. В один и тот же кластер относятся наблюдения, принадлежащие одному и тому же объекту.
3. Наблюдение может принадлежать только к одному классу (или кластеру).

### 4.2 Данные

Проверка эффективности применения рассмотренных методов машинного обучения производилась параллельно на двух наборах данных: вторая редакция данных наблюдений Gaia (data release 2) и данные из соревнования PLAsTiCC. Они отличаются по способу сбора данных, признакам, описывающим каждое наблюдения и количеству данных.

Gaia – это космический телескоп Европейского космического агентства, миссия которого состоит в построении трехмерной карты нашей Галактики, выявления ее состава, изучения ее эволюции [8]. На данный момент официально опубликованы две редакции набора данных. Вторая содержит результаты, основанные на наблюдениях, собранных в течение первых 22 месяцев с июля 2014 года. Она охватывает астрометрию, фотометрию, лучевые скорости, астрофизические параметры и многое другое. В рамках данной работы в качестве данных из него были выбраны две отдельных таблицы: данные об источниках (gaia_source) и данные кривых блеска (light_curves). Этот каталог включает около 100 различных признаков [7], [8], основными из которых являются: координаты (прямое восхождение и склонение), правильное движение источника, параллакс, величина потока в трех полосах пропускания (видимый зеленый, интегрированный синий, интегрированный красный). Также, каталог содержит большое количество дополнительной информации, такой как примерный радиус наблюдаемого объекта, эпоха при измерении или случайный индекс, используемый для выбора рассматриваемых подмножеств.

The Photometric LSST Astronomical Time Series Classification Challenge (PLAsTiCC) — это открытое соревнование по классификации смоделированных астрономических временных рядов при подготовке к наблюдениям с Большого синоптического телескопа [10]. Данные представляют собой временные ряды, содержащие информацию об излучении объектов в 6 фильтрах (ugrizy). В рамках нашей задачи был взят один из наборов соревнования, содержащий около 1,4 миллиона наблюдений для 7848 астрономических объектов [9]. Признаки в данном наборе [9], [10] делятся на две таблицы: постоянные, описывающие характеристики конкретного объекта независимо от наблюдений (например, координаты, красное смещение источника), и переменные, относящиеся к конкретному наблюдению (например, информация о спектре и величине потока в 6 различных спектрах (ugrizy)).

### 4.3 Предобработка данных

Данные каталога Gaia содержит большое количество признаков (около 100), а данные PLAsTiCC смоделированы и были созданы для другого типа задач.
Для того, чтобы подготовить их к данной работе, необходимо выполнить предобработку, состоящую из нескольких этапов.

**Генерация шума для постоянных признаков.** Этот этап необходим, так как в заданных каталогах значения наиболее важных параметров, которые нельзя не брать во внимание, как, например, координаты или красное смещение источника, были заданы постоянными для каждого объекта. В реальных же данных при получении каждого наблюдения объекта всегда существует некоторая погрешность измерения. Поэтому перед выполнением следующих этапов предобработки для каждого наблюдения таким признакам случайным образом прибавлялся шум в диапазоне [-1.5, 1.5].

**Отбор признаков.** Необходимо провести отбор признаков, чтобы выявить отрицательно влияющие на качество работы алгоритмов. Этот этап состоит из нескольких шагов:

1 Удаление всех постоянных признаков. Очевидно, что дополнительные признаки, имеющие постоянное значение для всех наблюдений одного и того же объекта, могут сразу указать, к какому источнику относится данное наблюдение. Например, такие данные, как модуль расстояния от объекта до измерительного прибора в PLAsTiCC или идентификаторы в Gaia.
2 Удаление неподходящих задаче признаков. Данный шаг больше относится к признакам из каталога Gaia, так как в нем есть те, что напрямую связаны с подсчетом информации из наблюдений (например, количество интегрированная средняя величина потока из определенной полосы).
3 Отбор признаков с использованием моделей. Для этого подхода необходимо использовать уже обученную модель машинного обучения, которая может показывать степень влияние признаков на ее работу. В рамках эксперимента использовался алгоритм случайного дерева. После обучения на данных он возвращает для всех признаков веса, показывающие их влияние на работу алгоритма. Этот же шаг позволил выявить признаки, на основе которых любой алгоритм для данной задачи дает 100% показатель точности. Так, например, случайный индекс имел столько же влияния, сколько и остальные признаки при классификации наблюдений из данных Gaia (Рис. 1).

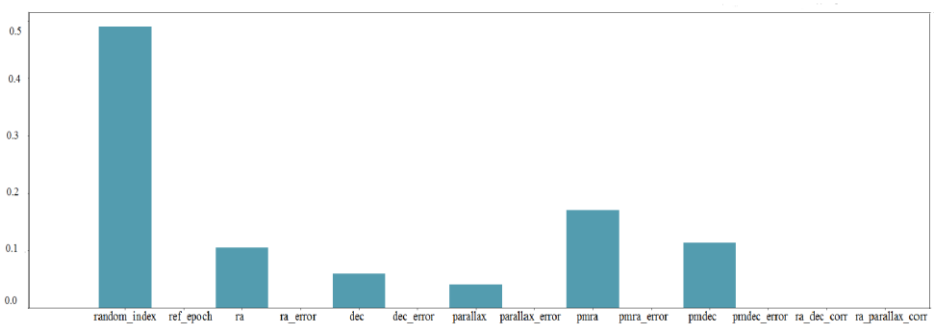


**Рис. 1.** Влияние признаков данных Gaia на результаты работа решающего дерева.

Отбор признаков с использованием моделей реализован уже после формирования обучающей и тренировочной выборок, во время обучения моделей.
В результате предобработки данных в используемом каталоге PLAsTiCC из 14 признаков осталось 9, а в данных Gaia осталось 22 признака из порядка 100.

### 4.4 Формирование обучающего и валидационного наборов данных

Прежде чем приступить к обучению модели и сравнению результатов работы различных алгоритмов в рамках задачи классификации, необходимо разделить данные на обучающую и валидационную выборку. В рамках первого эксперимента из каждого набора использовалось 100000 наблюдений. Эти наблюдения делились случайным образом в соотношении 7:3. После того как были получены высокие результаты качества работы моделей, было решено разделить исходные данные в соотношении 2:8 и обучить модели заново.

Однако такое формирование наборов данных оптимально не во всех случаях. Например, при таком разделении данных Gaia многослойный персептрон и машина опорных векторов выдавали очень низкие результаты. Это было связано с одним из важных признаков – величиной потока. В таблице наблюдений Gaia диапазон этого признака для одного и того же объекта может быть очень широким от 5000 до 11000. Это связано с тем, что таблица наблюдений хранит значения для трех типов полосы пропускания, а также имеет высокий уровень шума. Чтобы решить эту проблему, для данных Gaia специальным образом формировались тренировочный и валидационный наборы (также в соотношение 2:8): для каждого объекта в обучающий набор было отобрано по 7-8 наблюдений для каждого типа полосы пропускания, остальные данные формулировали валидационную выборку. После обучения на таком наборе многослойные персептрон и машина опорных векторов показали качество на 15% выше, чем при обучении на случайно созданном наборе. Однако полученный результат все еще намного ниже, чем результаты работы других алгоритмов. Это может быть связано с силой шумов при измерении потока. Решение этой проблемы будет рассмотрено в следующей работе.

### 4.5 Обучение модели и полученные результаты

Как было описано ранее, в рамках эксперимента для обучения и проверки качества работы методов машинного обучения параллельно было выбрано по 10000 наблюдений из каждого каталога.

Таким образом, на данном этапе рассматривался 31 астрономический объект из каталога PLAsTiCC, содержащий от 100 до 350 наблюдений каждый. Причем для большинства объектов количество наблюдений оказалось примерно одинаково. У Gaia широкий разброс наблюдений по объектам в рамках выбранных данных: от 45 до 165, поэтому и количество объектов, а значит и классов (кластеров), больше -103.

Итак, у нас есть два набора данных: один содержит 31 объект с 100- 350 наблюдениями, содержащими 9 признаков; и второй – 103 объекта с 45-165 наблюдениями, содержащими 22 признака. Второй набор более близок к реальности и кажется сложнее для обучения модели. Каждый каталог был предобработан и разделен на обучающую и валидационную выборки в соответствии с рассмотренными выше этапами. После этого к ним были применены выбранные методы классификации и кластеризации.

Для сравнения результатов обученных моделей машинного обучения было выбрано несколько показателей качества. В рамках классификации используются: accuracy (доля правильных ответов), precision (точность), recall (полнота), $F_1$-score ($F_1$-мера), потеря Хэмминга, коэффициент корреляции Мэтьюса. Для проверки качества кластеризации использовались: completeness (полнота), индекс Рэнда и V-мера.

**Полученные результаты.** Для выбора наилучшего метода машинного обучения в рамках данного эксперимента использовались несколько подходов. Для задачи классификации изначально проводилась перекрестная проверка, как и в статье [4], результаты этого этапа отражены в таблице 2. Далее данные делились на обучающую и валидационную выборку - на первой модель обучалась, на второй считались выбранные показатели качества. Кластеризация проводилась на всех 10000 наблюдений сразу, а далее сравнивалась с набором истинных меток с помощью выбранных показателей качества.

Результаты перекрестной проверки на данных PLAsTiCC отражены в таблице 2. Исходя из них, лучше всего себя показали: решающее дерево и многослойный персептрон, доля правильных ответов при классификации данными методами превысила 98%. Аналогичным образом все используемые алгоритмы показали себя при разделении данных в отношении 7:3. Чтобы проверить корректность полученных результатов, было решено применить их к обучающей и валидационной выборкам, сформированных в соотношении 2:8. На этом этапе были получены схожие результаты, то есть уменьшение в 3.5 раза обучающего набора не помешало алгоритмам качественно классифицировать данные (с точностью более 80%).

**Таблица 2.** Результаты перекрестной проверки классификаторов на данных PLAsTiCC.

| Показатель качества | Классификатор | Точность (%) |
|---|---|---|
| hamming_loss | KNeighborsClassifier | 5.1 – 6.9 |
| | **DecisionTreeClassifier** | **0 – 0.01** |
| | **MLPClassifier** | **0.6 – 1** |
| | SVC | 5.8 – 7 |
| accuracy_score | KNeighborsClassifier | 92.86 – 94.72 |
| | **DecisionTreeClassifier** | **99.94 – 99.99** |
| | **MLPClassifier** | **98.9 – 99.4** |
| | SVC | 92.8 – 94.2 |
| f1_score | KNeighborsClassifier | 91.58 – 93.6 |
| | **DecisionTreeClassifier** | **99.72 – 99.99** |
| | **MLPClassifier** | **98.86 – 99.45** |
| | SVC | 92.4 – 93.88 |

Агломеративная кластеризация, как и ожидалось, показала себя лучше остальных. Низкий уровень показателя индекса Рэнда, а также средний уровень V-меры связан с низким уровнем однородности, т.е. данные PLAsTiCC кластеризуются следующим образом: почти все наблюдения одного и того же объекта попадают в один кластер, но в этот же кластер попадает много наблюдений другого источника. Это связано с близким расположением этих объектов. Решение этой проблемы будет рассматриваться в следующих работах.

В результате перекрестной проверки на данных Gaia большинство алгоритмов показали результат хуже, чем на данных PLAsTiCC. Это связано со значением величины потока и смежными с ним признаками. Чтобы в этом удостовериться была проведены следующие тесты:

1 Формировались два набора данных случайным образом, как описано ранее, но без признака величины потока. На первом наборе данных, где обучающая выборка состояла из 70% данных, все алгоритмы показали результаты выше 90%, что согласуется с результатами на данных PLAsTiCC. Далее алгоритмы заново создавались и обучались на тренировочном наборе, содержащим всего 20% наблюдений. В этом случае результаты снизились лишь на 5%, т.е. все методы смогли качественно классифицировать заданные наблюдения.
2 Тренировочные и валидационный наборы формировались случайным образом с сохранением величины потока в соотношении 2:8. В данном случае решающее дерево хорошо справилось с поставленной задачей, получив значения показателей около 90%. Результаты алгоритма К-ближайших соседей снизились до 50%. Доля правильных ответов метода опорных векторов – 7%, f-мера – 15%. А вот многослойный персептрон показал себя хуже всех: доля правильных ответов меньше 2%.

Как было описано ранее, это связано с тем, что таблица наблюдений в Gaia хранит для каждого объекта значения величины потока для трех типов полосы пропускания и имеет высокий уровень шума, поэтому данный параметр может варьироваться от 4000 до 12000 для одного и того же объекта. Чтобы бороться с этой проблемой был вручную создан свой каталог. И уже на нем проверялось качество работы алгоритмов.

На данном этапе показатели метода опорных векторов улучшились следующим образом: доля правильных ответов и полнота показали результат около 25%, $F_1$-мера – 41%. Как видно, данный классификатор правильно распознал некоторый набор объектов, но соотносил к ним не только их наблюдения, но и наблюдения других объектов. Этот результат все еще является недостаточным для поставленной задачи. Возможно, это связано с силой шумов при измерении потока, но решение данной проблемы будет обсуждаться в следующей работе. Таким образом, на данных Gaia в задаче классификации лучше всего себя показали решающее дерево и алгоритм К-ближайших соседей.

Алгоритмы кластеризации в данном случае повели себя аналогичным образом: из-за величины потока все наблюдения почти соотносились к одному кластеру. Но при удалении данного признака, результат становился аналогичным резуль-

тату при работе с данными PLAsTiCC (чуть лучше показал себя алгоритм К-средних - он смог лучше разделить наблюдения разных объектов между собой, т.е. кластеры содержали в процентном соотношении больше наблюдений одного и того же источника).

Таким образом, эксперименты показали, что методы машинного обучения могут быть применены к задаче кросс-идентификации объектов и даже дать хороший результат. Одним из таких алгоритмов является решающее дерево, которое в проведенных экспериментах показало хорошие результаты. Кроме того, оно не требует ни нормализации, ни масштабирования данных, а его реализация проста для понимания. Оно спокойно работает и при отсутствии некоторых значений.

Однако структура данных и их качество все еще влияют на работу алгоритмов и требуют более высокого уровня этапа предобработки.

## 5 Заключение

В рамках данной статьи был проведен анализ существующих работ, связанных с применением методов машинного обучения в задачи кросс-идентификации астрономических объектов и их наблюдений. На их основе были выбраны алгоритмы, подходящие для поставленной задачи.

Выбранные алгоритмы классификации и кластеризации были протестированы на двух наборах (PLAsTiCC и Gaia). Лучший результат показали алгоритмы k-ближайших соседей и решающее дерево. Остальные оказались более зависимы к качеству и структуре собранных данных и требуют большего уровня предобработки. Решение данной проблемы будет рассмотрено в последующей работе.

В будущем необходимо будет решить проблемы, связанные с влиянием признака величины потока в данных Gaia, формированием обучающего и валидационного наборов. Также в последующей работе будет рассмотрено применение к данной задачи таких алгоритмов, как: XGBoost, LightGBM и RandomForest. А вместо MLPClassifier будет построена полноценная нейронная сеть с использованием библиотеки keras. Кроме того, будет проведена работа по исследованию применения метрик расстояния в задаче кластеризации с целью повышения качества работы рассматриваемых алгоритмов.

Работа выполняется в рамках магистерской диссертации на факультете Вычислительной математики и кибернетики МГУ им. М.В. Ломоносова под научным руководством Д.О. Брюхова, старшего научного сотрудника Федерального исследовательского центра «Информатика и управление» Российской Академии Наук.

### Литература

1. Clotet, M., Castaneda, J., Torra, J. , et al.: Cross-matching algorithm for the intermediate data updating system in Gaia (2016).
2. Lindegren, L.: Cross-matching Gaia objects (2015).
3. Torra, F., Clotet, M., Gonzalez-Vidal, J. J., et al.: Proper motion and other challenges in cross-matching Gaia observations (2018).

4. Rohde, D. J., Drinkwater, M. J.: Applying machine learning to catalogue matching in astrophysics (2005).
5. Малков, О.Ю., Длужневская, О.Б., Кайгородов, П.В., Ковалева, Д.А., Скворцов, Н.А.: Проблемы обозначения и кросс-идентификации кратных объектов в астрономии (2015).
6. Угадяров, Л. А.: Реализация методов распределенной обработки астрономических изображений (2019).
7. Официальный сайт архива данных Gaia: https://gea.esac.esa.int/archive/ (10.05.2020)
8. Lindegren, L.: Gaia Data Release 2: The astrometric solution (2018).
9. Allam, T. Jr., Anita Bahmanyar, A., Biswas, R., et al.: The Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC): Data set (2018).
10. Официальный сайт соревнования PLAsTiCC: https://www.kaggle.com/c/PLAsTiCC-2018 (14.05.2020)
11. Документация библиотеки sklearn: https://scikit-learn.org/stable/index.html (20.05.2020)