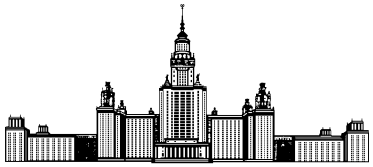
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

«Кластеризация галактик из каталога SDSS»

Выполнил:

студент 3 курса 317 группы

Демин Георгий Александрович

Научный руководитель:

д.ф-м.н., профессор РАН

Дьяконов Александр Геннадьевич

Москва, 2019

Содержание

1 Введение **3**

1.1 Определения и обозначения 4

1.2 Обзор литературы 5

2 Новые подходы и результаты **5**

3 Вычислительные эксперименты **5**

3.1 Исходные данные и условия эксперимента 6

3.2 Результаты эксперимента 6

3.3 Обсуждение и выводы 6

4 Заключение **6**

Список литературы **8**

Аннотация

Выделение скоплений галактик - важная задача современной астрофизики. Изучая распределение кластеров галактик (и сами кластеры) в космосе, можно исследовать вселенную и процессы, происходящие на разных этапах её развития.

В работе приведён обзор некоторых алгоритмов кластеризации, а также результаты их применения к решению задач кластеризации галактик.

1 Введение

Задача кластеризации в машинном обучении - это задача разбиения выборки объетов на некоторые множества, называющиеся кластерами, при этом одному множеству должны принадлежать в некотором роде похожие объекты, а объекты из разных кластеров - существенно отличаться. Входными данными для задачи кластеризации могут быть как объекты и их признаковое описание, так и матрица расстояний (или схожести) между объектами. Обычно эта задача является промежуточным этапом построения модели в машинном обучении. Она может использоваться для добавления к объектам признаков, уменьшения объёма данных путём объеденения объектов, попавших в один кластер. Также целью может служить - понимание структуры данных и их распределения.

Особенностью этой задачи является то, что её решение принципиально неоднозначно: нет точной постановки задачи, существует много критериев качества кластеризации, результат решения задачи сильно зависит от выбранной метрики (или схожести близости), качество кластеризации определяется во многом спецификой задачи (один и тот же результат может восприниматься как хороший или плохой в зависимости от того, для чего именно мы кластеризуем данные).

Для решения задачи могут применяться различные алгоритмы в зависимости от того, какой тип кластеров ожидается. На основании этого задачу кластеризации разделяют на (ссылка!):

- **Иерархическая** или **плоская**. Под иерархической структурой кластеров понимает такое их устройство, что некоторые мелкие кластеры могу быть объеденены в бóльшие
- мультиномиальная регрессия
- методы многоклассовой классификации one-vs-all и all-vs-all
- градиентный и стохастический градиентый спуск
- некоторые методы обработки текстов

Во введении рассказывается, где возникает данная задача, и почему её решение так важно. Вводится на неформальном уровне минимум терминов, необходимый для

понимания постановки задачи. Приводится краткий анализ источников информации (литературный обзор): как эту задачу решали до сих пор, в чем недостаток этих решений, и что нового предлагает автор. Формулируются цели исследования. В конце введения даётся краткое содержание работы по разделам; при этом отмечается, какие подходы, методы, алгоритмы предлагаются автором впервые. При упоминании ключевых разделов кратко формулируются основные результаты и наиболее важные выводы.

Цель введения: дать достаточно полное представление о выполненном исследовании и полученных результатах, понятное широкому кругу специалистов. Большинство читателей прочтут именно введение и, быть может, заключение. Во введении автор решает сложную оптимизационную проблему: как сообщить только самое важное, потратив минимум времени читателя, да так, чтобы максимум читателей поняли, о чём вообще идёт речь.

Введение лучше писать напоследок, так как в ходе работы обычно происходит переосмысление постановки задачи. Если же введение писать, когда работа еще не готова, задача усложняется вдвойне. В конце обычно приходит понимание, что всё получилось совсем не так, как планировалось в начале, и исходный вариант введения всё равно придётся переписывать. Кстати, к таким «потерям» надо относиться спокойно — в хорошей работе почти каждый абзац многократно переделывается до неузнаваемости.

Введение имеет много общего с текстом доклада на защите, поэтому имеет смысл готовить их одновременно.

1.1 Определения и обозначения

Формальная постановка задачи. Для известных понятий желательно придерживаться стандартных обозначений. Общепринятые термины вводятся словом «называется». Термины, придуманные самим автором, вводятся словами «назовём» или «будем называть». Обычно этот раздел заканчивается формальной постановкой задачи. Именно с этого раздела стоит начинать писать работу.

1.2 Обзор литературы

Лучше, чтобы название этого подраздела было содержательным, например, общепринятым названием задачи, проблемы или метода, рассматриваемого в данной работе.

Перечисляются подходы, методы, факты, на которые существенно опирается данная работа, но которые могут быть не известны широкому кругу читателей. Здесь ссылки на литературу обязательны. Теоремы только формулируются, но не доказываются.

Данный раздел преследует две цели. Во-первых, сделать работу самодостаточной — дать необходимый минимум информации тем читателям, которые не очень хорошо ориентируются в теме, но желают поближе познакомиться именно с данной работой. Во-вторых, облегчить сопоставление полученных автором результатов с ранее известными.

2 Новые подходы и результаты

Название этого раздела обязательно надо заменить на содержательное. В этом разделе, как правило, много подразделов.

В дипломной работе не стоит делать более двух уровней, достаточно разделов и подразделов. Будете писать диссертацию или монографию — сделаете три уровня.

3 Вычислительные эксперименты

Цель данного раздела: продемонстрировать, что предложенная теория работает на практике; показать границы её применимости; рассказать о новых экспериментальных фактах.

Чисто теоретические работы могут вообще не содержать раздела экспериментов (не работает, ну и не надо — зато теория красивая). Кстати, теоретики имеют право не догадываться, где, кому и когда их теории пригодятся.

3.1 Исходные данные и условия эксперимента

Описывается прикладная задача, параметры анализируемых данных (например, сколько объектов, сколько признаков, каких они типов), параметры эксперимента (например, как производился скользящий контроль).

3.2 Результаты эксперимента

Результаты экспериментов представляются в виде таблиц и графиков. Объясняется точный смысл всех обозначений на графиках, строк и столбцов в таблицах.

3.3 Обсуждение и выводы

Приводятся выводы: в какой степени результаты экспериментов согласуются с теорией? Достигнут ли желаемый результат? Обнаружены ли какие-либо факты, не нашедшие объяснения, и которые нельзя списать на «грязный» эксперимент?

Обсуждаются основные отличия предложенных методов от известных ранее. В чем их преимущества? Каковы границы их применимости? Какие проблемы удалось решить, а какие остались открытыми? Какие возникли новые постановки задач?

4 Заключение

В квалификационных работах последний раздел нужен для того, чтобы конспективно перечислить основные результаты, полученные лично автором.

Результатами, в частности, являются:

- Предложен новый подход к...
- Разработан новый метод..., позволяющий...
- Доказан ряд теорем, подтверждающих (опровергающих), что...
- Проведены вычислительные эксперименты..., которые подтвердили / опровергли / привели к новым постановкам задач.

Цель данного раздела: доказать квалификацию автора. Даже беглого взгляда на заключение должно быть достаточно, чтобы стало ясно: автору удалось решить

актуальную, трудную, ранее не решённую задачу, предложенные автором решения обоснованы и проверены.

Иногда в Заключении приводится список направлений дальнейших исследований.

Список литературы необходим в любой научной публикации. В дипломной работе он обязателен. Дурным тоном считается: ссылаться на работы только одного-двух авторов (например, себя или шефа); ссылаться на слишком малое число работ; ссылаться только на очень старые работы; ссылаться на работы, которых автор ни разу не видел; ссылаться на работы, которые не упоминаются в тексте или которые не имеют отношения к данному тексту.