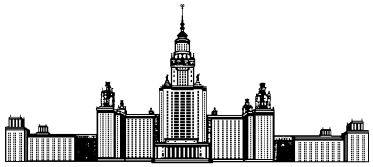
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

«Задача кластеризации сильно зашумлённых данных с неоднородной плотностью »

Выполнил:

студент 3 курса 317 группы

Демин Георгий Александрович

Научный руководитель:

д.ф-м.н., профессор РАН

Дьяконов Александр Геннадьевич

Москва, 2019

Содержание

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.

Не готовы аннотация, введения, эксперименты и вывод

1 Введение. Кластеризация

Оно пока не написано, возьму часть из постановки задачи

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.

2 Постановка задачи

2.1 Задача кластеризации в машинном обучении

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 структура кластеров ожидается. На основании этого задачу кластеризации разделяют на (ссылка!):

- **Иерархическая** или **плоская**. Под иерархической структурой кластеров понимается такое их устройство, что некоторые мелкие кластеры могу быть объеденены в более крупные, которые в свою очередь также могут быть объеденены на более высоком уровне иерархии. Плоская кластеризация подразумевает, что некоторое множество выделенных кластеров не может само образовывать кластер.

- **Исключающая**, **перекрывающая** или **нечёткая**. Исключающая кластеризация означает, что объект может быть отнесён только к одному классу, перекрывающая - к нескольким. Нечёткая кластеризация является частным случа-

ем перекрывающей и для каждого объекта определяет вероятность вхождения в каждый кластер (другое её название - **вероятностная**).

- **Полная** или **частичная**. При полной кластеризации каждый объект обязательно относится к кому-либо кластеру, а при частичной объект может являться шумом или выбросом и не должен принадлежать кластеру.

- **Прототипная**, **графовая** или **плотностная**. При прототипной кластеризации кластер определяется как множество объектов, похожих на некий прототип. Графовая модель — модель, при которой данные представлены в виде графа и кластером будет являться некое подмножество вершин, которые имеют большое количесвто связей или образуют отдельную связную компоненту. Плотностная кластеризация же означает, что кластером будут считаться объекты, плотность (при каком-то заданном распределении) которых больше средней плотности всей выборки.

В данной работе мы будем решать плоскую исключающую плотностную задачу кластеризации. Кроме того, ожидаем, что большая часть данных относится к шуму

2.2 Формальная постановка задачи

Приведём формальную постановку задачи: имеется $X \in \mathbb{R}^{N \times D}$ - матрица объекты-признаки, $D \ll N$. $S = \{0, 1, ..., K\}$ - множество номеров кластеров, K (число кастеров) априорно неизвестно точно. Требуется найти отображение $y(x) : X \to Y$, где $Y_i \in S$ и является номером кластера, которому принадлежит X_i. $y(x_0) = 0$ означает то, что объект x_0 является шумовым и не принадлжеит ни одному кластеру. Кроме того, мы предполагаем, что

- Кластеризация производится на основе плотности

- Число объектов, не относящихся ни к какому кластеру велико.

$$\frac{|\{x|y(x) = 0\}|}{|X|} \geqslant 0.5 \tag{1}$$

.

- Имеется некое априорное представление о структуре кластеров

$$\exists F_i(Y, K) \approx C_i, i = \overline{1, m} \tag{2}$$

3 Обзор существующих решений

3.1 Сдвиг среднего значения (mean shift)

Описание алгоритма mean shift

3.2 DBSCAN

Описание DBSCAN

Далее рассмотрим некоторые модификации алгоритма DBSCAN, на которые будет существенно опираться предложенный ниже подход

3.2.1 VDBSCAN

Для набора данных с варирующейся плотностью в статье [1] был предложен следующий подход: запустить алгоритм DBSCAN несколько раз с разными значениями радиуса. При каждой последующей итерации исключать точки, кластеризованные на предыдущей, - таким образом могут быть обнаружены кластеры с различной плотностью. В статье [2] описывается способ для определения параметров k и eps. Предлагается посчитать среднее расстояние между точками в датасете, затем для каждой точки посчитать какой по номеру ближайший сосед отстоит от неё на это расстояние и взять в качестве k - самый частый номер среди всех точек. Далее для каждой точки находится расстояние от неё до k-го ближайшего соседа, строится гистограмма таких расстояний и резкие скачки на ней будут считаться кандидатами на eps. У этого подхода есть несколько недостатков: во-первых, в нём слабо учитывается априорные знания о данных (только то, что они имеют варирующуюся плотность), во-вторых, он не подходит для датасета с большим количеством шума, так как гистограмма расстояний до k-го ближайшего соседа не позволит определить значения eps (она будет слишком гладкой из-за расстония до шумовых точек)

3.2.2 Дифференциальная эволюция для подбора параметров DBSCAN

Дифференциальная эволюция — это метод стохастической оптимизации(см. [4]). Его идея в том, что генерируется некоторое множество алгоритмов с разными параметрами - популяции (при этом используется один и тот же алгоритм и изменяются лишь его параметры), для каждого алгоритма вычисляется функция потерь. Алгоритмы, с "удачной" функцией потерь (то есть значение которой, оказалось низким), скрещиваются и получившийся алгоритм получает параметры, близкие к параметрам его предков. неудачные же алгоритмы с некоторой вероятностью мутируют, изменяя свои параметры. Несомненное преимущество этого метода в том, что он подходит для оптимизации вообще любых алгоритмов (однако вопрос о сходимости остаётся открытым). В [3] авторы используют этот подход, чтобы подобрать k и eps для алгоритма DBSCAN. Это довольно удачный способ для нашей постановки задачи, так как все 2 могут быть учтены в функции потерь и параметры будут подобраны строго автоматически. Однако в этом методе тяжело учесть то, что данные неоднородные.

Для алгоритма DBSCAN есть множество усовершенствований, однако каждое из них имеет собственные недостатки, не позволяющие решить исследуемую задачу достаточно хорошо. Ниже будет представлен метод, опирающийся на приведённые выше модификации DBSCAN.

4 Предлагаемый подход

В основе предлагаемого подхода лежит одна простая идея: для областей с разной плотностью нужно использовать алгоритм с разными параметрами. Сначала опишем метод, а затем приведём его формальное описание

1. Разбиваем признаковое пространство на подпространства с более-менее однородным распределением признаков(для этого можно построить распределение признаков и делить некоторые из них)

2. Определяем некоторую функцию потерь, зависящую от (2), с помощью неё на каждом подпространстве находим параметры алгоритма DBSCAN, при этом

слишком маленькие по размеру выделенные кластеры относим к шуму, чтобы учесть (1)

3. Выделяем из подпространств зоны, на которых работа алгоритма нас не устраивает (это можно сделать на основе функции потерь или отдельных условий 2) И повторяем шаги независимо в каждой такой зоне.

Algorithm 1 DBSAN for variety density

function OPTIMIZEDBSCAN(X, lossFunc)

$X = newint[5]$

$subareas \leftarrow$ GETSUBAREAS(X)

end function

function TUNEONSUBAREAS($subareas$)

for $X_i \in subareas$ **do**

eps, k $\leftarrow$ OPTIMIZEDBSCAN(X_i, LossFunc)

labels $\leftarrow$ labels $\cup$ DBSCAN(X_i, eps, k)

end for

end function

5 Набор данных. Каталог SDSS

Продемонстрируем предлагаемый алгоритмна задаче выделения галактик по каталогу SDSS(ссылка?). Этот каталог содержит сферические координаты и некоторые физические величины более чем 3 миллионах галактик, наблюдаемых в северном полушарии Земли. **Расстояние** до галактики определяется с помощью измерения красного смещения электромагнитных волн, излучаемых ею (ссылка!), близкими считаются галактики, имеющие красное смещение меньше 0.3 (всюду далее под расстоянием до галактики мы будем подразумевать красное смещение этой галактики). Особенностью этих данных является то, что для галактик сильно удалённых от Земли погрешность вычислений координат достигает больших значений - это не позволяет их анализировать. **Кластером** (скоплением или глобуларом) называется набор галактик, расположенных близко друг к другу (более точного определения дать не удаётся, так как объективные характеристики, по которым можно было бы определить что является кластером, а что нет, пока в астрофизике не определены). Считается также, что галактика принадлежит только одному кластеру или не принадлежит никакому и что примерный размер кластера порядка 0.001 (в единицах красного смещения), а расстояние между ними 0.004.

6 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.

6.1 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).

6.2 Результаты эксперимента

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

6.3 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?

7 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.

8 Литература

Список литературы

[1] Peng Liu, Dong Zhou, Naijun Wu, "VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise" // 2007 International Conference on Service Systems and Service Management

[2] A. Özekes, F. Ozge Ozkok et al, "AutoVDBSCAN: An Automatic and Level-Wise Varied-Density Based Anomaly Detection Algorithm" // 2018 Conference: 7th International Conference on Advanced Technologies (ICAT'18)

[3] A. Karami, R. Johansson, "Choosing DBSCAN Parameters Automatically using Differential Evolution" // 2014 International Journal of Computer Applications (0975 8887) Volume 91 - No.

[4] R. M. Storn, K. Price, "Differential Evolution: A Simple and Efficient Adaptive Scheme for Global Optimization Over Continuous Spaces" // 1995 Journal of Global Optimization

[5] q

[6] q

[7] q

[8] q

[9] q

[10] q