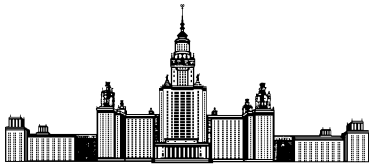
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

«Задача кластеризации 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, часть экспериментов и вывод.

1 Введение

Задача кластеризации в машинном обучение - типичная задача обучения без учителя: есть некоторые критерии, которые сигнализирует о том хорошо решена задача или нет, но точной информации нет. 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. Более подробно задача кластеризации описана в разделе "постановка задачи".

В данной работе рассматривается задача кластеризация на данных, о которых известно, что большая их часть в кластеры не входит и является шумом. Кроме того, известно, что данные неоднородны и их плотность варируется. Предполагается также, что имеется некоторая априорная информация о распределении кластеров. Такая задача может возникнуть при распознавании образов, анализе биологических или физических систем.

Предлагается алгоритм, основанный на DBSCAN, учитывающий некоторые априорные представления о данных и адаптирующийся, способный настраиваться на варьирующуюся плотность. Алгоритм используется для решения задачи кластеризации галактик из каталога SDSS. Результаты его работы сравниваются с результатами, полученными с помощью "классических" алгоритмов (mean shift, классический DBSCAN)

Опишем более подробно задачу кластеризации и её типы.

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.

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 ([1]):

- **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 случаем 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 [2] 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 [3] 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. [5]). 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 [4] 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 по-

добраны 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 TuneOnSubareas(*subareas*)

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

    eps, k $\leftarrow$ OptimizeDBSCAN(X_i, LossFunc1)

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

  end for

  return *labels*

end function

$subareas \leftarrow$ GetSubareas(X)

$labels \leftarrow \{\}$

$labels\_first\_step \leftarrow$ TuneOnSubareas($subareas$)

$bad\_labels \leftarrow$ FindBadLabels(labels, LossFunc2)

$bad\_subareas \leftarrow subareas[bad\_labels]$

$labels\_first\_step \leftarrow labels\_first\_step \setminus bad\_labels$

$labels\_result \leftarrow labels \cup$ TuneOnSubareas($bad\_subareas$)

return $labels\_result$

Функция GetSubareas неким образом разбивает выборку X на подобасти LossFunc1, LossFunc2 неким образом отражают апрорную информацию и штрафуют за нарушение (2) OptimizeDBSCAN подбирает параметры для алгоритма DBSCAN на подвыборке, используя дифференциальную эволюцию, а DBSCAN применяет алгоритм с подобранными параметрами. Наконец, FindBadLabels определяет какие данные были кластеризованы плохо (в следствие различия в плотности или чего-то другого).

Далее покажем, как все можно составить функции потерь, используя априорные знания на реальных данных.

5 Набор данных. Каталог SDSS

Продемонстрируем предлагаемый алгоритм на задаче выделения галактик по каталогу SDSS([6]). 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, 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 либо принадлежит только одному кластеру, либо не принадлежит никакому (таких галактик большинство и мы будем называть их шумовыми). Ожидается, что кластеры будут сферической или элипсовидной формы

6 Эксперименты

6.1 Априорные представления и их формализация

В качестве априорных представлений о струтуре кластеров мы возьмём результаты работы Дж. Эйбелла [7]. Американский астроном вручную проанализировал фотопластины северной части звёздного неба и выделил 2712 кластеров. Так как это было сделано по фотометрическим данным, нельзя сказать, что координаты выделенных кластеров точны. Мы будем использовать 2 факта из работы физика: число кластеров приближённо известно и равно 1570[1], характерный размер кластера примерно 125-175 галактик. Обозначим за $N(k)$ количество объектов в кластере с номером k, и фоормализуем априорную информацию таким образом

$$F_1(Y,K) = F_1(|Y|,K) = |K - 1572 \cdot \frac{|Y|}{|X|}| - 50 \leq 0 \tag{3}$$

$$F_2(Y,K) = F_2(N(1),\ldots,N(K),K) = \frac{1}{K}\sum_{k=1}^{K}(|N(k) - 150| - 50) \leq 0 \tag{4}$$

[1]Каталог SDSS покрывает не все области видимого неба, поэтому ожидается меньшее число кластеров

Неравенства (3) и (4) не стоит понимать строго: это лишь наши предположения, они могут нарушаться. Исходя их этого построим функцию потерь для оптимизации DBSCAN. Будем штрафовать алгоритм, если он выделяет не то количество кластеров, которое нам нужно и если в кластерах оказывается намного больше или меньше галактик, чем предполагалось. Для этого хорошо подходит квадратичная функция.

$$f_1(K) = a_1 K^2 + b_1 K + c_1$$

$$f_2(N(i)) = a_2 N(i)^2 + b_2 N(i) + c_2$$

Коэффициенты определим из условий

$$\begin{cases} f_1|_{K=1572\cdot\frac{|Y|}{|X|}} = 0 \\ f_1|_{K=1572\cdot\frac{|Y|}{|X|}+50} = 1 \\ f_1|_{K=1572\cdot\frac{|Y|}{|X|}-50} = 2.5 \end{cases} \tag{5}$$

$$\begin{cases} f_2|_{N(i)=150} = 0 \\ f_2|_{N(i)=100} = 2 \\ f_2|_{N(i)=200} = 1 \end{cases} \tag{6}$$

Ясно, что получается 2 СЛАУ, коэффициенты которых легко находятся. Заметим, что мы полагаем квадратичную функцию равной нулю в точках с ожидаемым значением, и накладываем некоторый штраф в точках, удалённых на 50 от ожидаемого значения — этот отступ, как и штрафы, вообще говоря, можно выбирать по-другому. Введём, наконец, функции потерь.

LossFunc2 $= \mathfrak{L}_2(Y, K) = \underset{i}{arg}\{N(i) > T\}$

LossFunc1 $= \mathfrak{L}_1(Y, K) = f_1(K) + \frac{1}{K}\sum_{i=1}^{K\setminus\mathfrak{L}_2(Y,K)} f_2(N(i))$

Эти функции с одной стороны учитывают априорные условия и с другой учитывают то, что данные имеют неоднородною плотность. Действительно, мы настраиваем DBSCAN таким образом, чтобы кластеры получались характерного небольшого размера, но если какой-то кластер получается слишком большим, мы его не учитываем и не штрафуем за него, вместо этого считаем, что распределение объектов, входящих в него имеет большую плотность и требуется для них подобрать другие параметры.

Заметим, что мы никак не штрафуем количество галактик, которые определяются как шумовые, так как ограничения на количество объектов в каждом кластере и на количество самих кластеров по сути учитывают это.

возвращает номера кастеров, количество объектов в которых больше некоторого порога. Эта функция отражает условие неоднородности плотности данных: действительно, установив порог достаточно большим (например, в 10 раз больше ожидаемого числа кластеров), мы получим кластеры, в которых оказалось слишком много галактик. Ясно, что плотность распределения объектов, попавших в эти кластеры больше, чем средняя плотность по всей выборке, однако мы считаем, что это может

6.2 Применённые алгоритмы

Мы будем сравнивать результаты работы 3 основных алгоритмов

- Mean shift
- Mini Batch K-means (спецификация K- means, которая для пересчёта центра кластера на каждой итерации использует не все объекты, принадлежащие ему, а лишь некоторую их подвыборку, что очень ускоряет алгоритм)
- Предложенный алгоритм оптимизации DBSCAN с дополнительным подбором параметров на областях с высокой плотностью

6.3 Метрики качества

Для задачи кластеризации с неизвестным числом кластеров существуют несколько коэффициентов, в какой-то мере описывающих резульат

- Silhouette Coefficient
- Davies-Bouldin Index
- Calinski-Harabasz Index (Variance Ratio Criterion)

$$s(K) = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{|X| - K}{K - 1}$$

где B_k - матрица дисперсий центров кластеров, W_k - сумма матриц дисперсий всех кластеров

$$B_k = \sum_k N(k)(c_k - c)(c_k - c)^{\mathrm{T}}$$

$$W_k = \sum_{k=1}^{K} \sum_{x:y(x)=k} (x - c_k)(x - c_k)^{\mathrm{T}}$$

c_k - центр k-го кластера, c - центра всех объектов. Подразумевается, что метрика расстояний между объектами (и соответсвенно кластерами) - евклидова. Чем $s(K)$, тем лучше результат кластеризации. Дейсвительно максимизация $s(K)$ равносильна максимизации дисперсий распределения центров кластеров (то есть, чем более центры кластеров будут разбросаны, тем лучше) и в то же время минимизации дисперсий распределения в объектов в каждом кластере (чем компактнее расположены объекты, тем лучше).

Все эти 3 коэффициента дают лучшие значение для выпуклых сферических кластеров с евклидовой метрикой. Это подходит для нашей задачи.

Отметим то, что обысно DBSCAN не склонен выделять кластеры сферической формы, его главное преимущество состоит в том, что он способен распознавать кластеры производной формы.

6.4 Сравнения методов

	# subareas	silhouette	(C-H) / 10^3	(1− D-B)
Mean Shift	4	0.34	21	0.15
Mini Batch K-means	4	0.22	**49**	-0.1
optimized DBSCAN	5	0.11	12	0.23
optimized DBSCAN NH	5	**0.42**	41	**0.32**
optimized DBSCAN	4	0.13	17	0.23
optimized DBSCAN NH	4	**0.49**	**59**	**0.4**

Таблица 1: Сравнение метод кластеризации

В Таблице 1 приведены результаты работы различных методов. Колонка #subareas показывает насколько частей делилась исходная выборка; silhouette содер-

жит значения соответствующего коэффициента; колонка C-H/10^3 заполнена значениями Calinski-Harabasz Index, делёнными на 1000; (1-DB) - Davies–Bouldin index, вычтенные из 1 (так как лучшей классификации соответствует меньшее значение, применяем такое преобразование для согласованности с другими индексами). optimized DBSCAN - это оптимизированный алгоритм без дополнительного подбора параметров на областях, соответствующих большим кластерам; optimized DBSCAN NH - с дополнительным подбором и разбиением соответсвующих областей

6.5 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] M. Steinbach, V. Kumar, "Cluster Analysis: Basic Concepts and Algorithms", January 2005

[2] 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

[3] 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)

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

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

[6] SDSS Collaboration, "THE THIRTEENTH DATA RELEASE OF THE SLOAN DIGITAL SKY SURVEY: FIRST SPECTROSCOPIC DATA FROM THE SDSS-IV SURVEY MAPPING NEARBY GALAXIES AT APACHE POINT OBSERVATORY"

[7] Abell, G. O. (1958)."The distribution of rich clusters of galaxies. A catalogue of 2712 rich clusters found on the National Geographic Society Palomar Observatory Sky Survey". The Astrophysical Journal Supplement Series, 3,: 211—288.

[8] q