Мягкие цепи ближайших соседей

краткое содержание

М. Мрыхин

5 апреля 2019 г.

Определение 1. *Задача* ***агломеративной иерархической кластеризации*** *состоит в следующем: дано начальное множество элементов S и кластерная метрика $d : \{\, x, y \in 2^S \mid x \cap y = \emptyset \,\} \to \mathbb{R}$. Требуется построить двоичное дерево подмножеств S, являющееся результатом работы данного алгоритма: начать с множества $R = \{\, \{x\} \mid x \in S \,\}$; на каждом шаге взять два ближайших элемента R, заменить на их объединение, а в дереве сделать объединение их родителем; продолжать, пока $R \neq \{S\}$.*

Примечание: вообще кластерная метрика не обязана быть метрикой в общепринятом смысле, от неё требуется только симметричность. Однако далее рассматриваются метрики семейства L_p.

Определение 2. *Кластерная метрика d называется* ***сводимой****, если для любых трёх дизъюнктных подмножеств A, B и C $\quad d(A \cup B, C) \geq \max(d(A, C), d(B, C))$.*

Лемма 1. *Если метрика d сводима и не допускает равных расстояний между разными парами подмножеств, то исходный алгоритм и алгоритм, объединяющий вместо ближайшей пары соседей пару взаимно ближайших соседей, строят одно и то же дерево.*

Свойство выше называется **глобально-локальной эквивалентностью**.

Определение 3. *Алгоритм* ***цепи ближайших соседей (ЦБС)*** *работает следующим образом: если стек пуст, добавить в него произвольное подмножество из R; иначе, если ближайший сосед верхнего подмножества в стеке не является взаимным, добавить его в стек; иначе объединить верхнее подмножество в стеке с ближайшим соседом и удалить из стека; повторять, пока все подмножества не объединены в S.*

Но алгоритмы поиска ближайших соседей не настолько эффективны, как хотелось бы, так что обобщаем.

Определение 4. ***Динамическая ε-приблизительная структура k ближайших соседей (k-ПБС)*** *- это структура, поддерживающая множество точек P и запросы на добавление, удаление и ε-приблизительный поиск k ближайших соседей: по точке p найти k различных точек $p_1, p_2, \dots, p_k \in P, p_i \neq p$, так что $d(p, p_i) \leq (1+\varepsilon)d(p, p_i^*)$, где p_i^* - i-тая ближайшая точка к p.*

Лемма 2. *Для любых констант $\delta, k \in \mathbb{N}, p > 1$ и $\varepsilon > 0$ существует k-ПБС, обрабатывающая точки в $\mathbb{R}^\delta$ по метрике L_p со временем инициализации $O(n \log n)$ и временем операций $O(\log n)$.*

Примечание: эта лемма не доказывалась, она зацитирована из другой статьи.

Определение 5. ***Динамическая мягкая структура ближайших соседей (МБС)*** *- это структура, поддерживающая множество точек P и запросы на добавление, удаление и мягкий поиск ближайшего соседа: по точке p найти ближайшую к ней точку p_1^*, либо пару точек q, q', лежащих друг к другу ближе, чем p и p_1^*.*

Для удобства будем считать, что в первом случае (т.н. **жёсткий ответ**) искомая точка выдаётся в паре с запрашиваемой. Далее мы строим МБС-структуру из k-ПБС-структуры с подходящими k и ε:

Лемма 3. *Если k-ПБС-структура на поисковый запрос выдаёт k точек, отличных от p_1^*, то для каждого i $d(p, p_i) \leq (1+\varepsilon)^i d(p, p_1^*)$.*

Будем называть (ε, k) **корректными параметрами**, если среди любых k точек в сферической оболочке с внутренним радиусом 1 и внешним радиусом $(1+\varepsilon)^k$ найдётся пара на расстоянии меньше 1.

Лемма 4. *Для любого метрического пространства $(\mathbb{R}^\delta, L_p)$ существуют корректные параметры.*

Лемма 5. *Если (ε, k) - корректные параметры, то следующий алгоритм моделирует поисковый запрос МБС-структуру на базе k-ПБС-структуры: задать поисковый запрос относительно той же точки; выдать ближайшую пару из запрашиваемой и/или выданных в ответ на запрос точек.*

И из всего этого собирается

Теорема 1. *Для любых констант $\delta \in \mathbb{N}$ и $p > 1$ существует МБС-структура, обрабатывающая точки в $\mathbb{R}^\delta$ по метрике L_p со временем инициализации $O(n \log n)$ и временем операций $O(\log n)$.*

Определение 6. *Алгоритм* ***мягкой цепи ближайших соседей (МЦБС)*** *работает следующим образом: если стек пуст, добавить в него результат запроса относительно произвольного подмножества из R; иначе, если среди результатов запросов относительно верхней пары подмножеств есть хотя бы одна пара подмножеств на меньшем расстоянии, добавить в стек ближайшую из этих пар; иначе объединить верхнюю пару подмножеств в стеке и удалить из стека, а также удалить верхнюю пару ещё раз, если у неё был общий элемент с удалённой; повторять, пока все подмножества не объединены в S.*

Лемма 6. *Между итерациями алгоритма МЦБС все подмножества в стеке принадлежат R; более того, если одно и то же подмножество находится в двух парах, эти пары являются соседями в стеке.*

Теорема 2. *Алгоритм МЦБС работает корректно и за время $O(P(n) + nT(n))$, где $P(n)$ - время инициализации, а $T(n)$ - время операций.*

Следствие 1. *В случае, когда кластеры реализуются как точки в метрическом пространстве $(\mathbb{R}^\delta, L_p)$, алгоритм МЦБС работает за время $O(n \log n)$.*