УДК 004.8

МАСШТАБИРУЕМАЯ АППРОКСИМАЦИЯ UPGMA ДЛЯ ЕВКЛИДОВЫХ ДАННЫХ

Наумов С.С. (Университет ИТМО)
Научный руководитель – к.ф.-м.н., доцент Фильченков А.А
(Университет ИТМО)

Многие широко используемые методы анализа данных, применяемые в исследовательских областях, плохо работают при использовании больших наборов данных. Один из популярных методов, агломерационная иерархическая кластеризация UPGMA имеет квадратичную сложность и плохо масштабируется на большое количество объектов. В этом докладе мы предлагаем использовать local sensitive hashing в сочетании с рандомизированным ядерным методом для получения быстрой пошаговой аппроксимации UPGMA с теоретической гарантией оптимизации Mosgly-Wang objective.
Ключевые слова: иерархическая кластеризация, аппроксимационный алгоритм, UPGMA, LSH, kernel trick

Введение. Иерархическая кластеризация является универсальным инструментом для поиска выразительного представления данных, анализ данных и визуализации. Иерархическая кластеризация может обнаружить древовидную структуру представления, которые используются в таких задачах, как персонализация и рекомендации. В эпоху нейронных сетей большие и высокоразмерные данных становится широко используемым, поэтому возникает необходимость в улучшение и обобщение существующих алгоритмов с целью оптимизации времени и памяти.
Наиболее ярким иллюстративным применением иерархической кластеризации является филогенетика: дано попарно сходства или расстояния между различными видами, необходимо рекурсивно разделять группы в меньшие кластеры. Таким образом получается дерево в котором каждый уровень соответствует типу, роду, царству и тд, а листы — это исходные объекты. Такое дерево называется дендрограммой и может использоваться для формирования любого числа кластеры, поиска закономерностей, визуализации.
Unweighted Pair-Group Method Using Arithmetic Averages или сокращенно UPGMA — это один из самых популярных методов агломерационной иерархической кластеризации. На каждом шаге алгоритм ищет два самых ближайших кластера по среднему расстоянию и объединяет их. На первом шаге каждый объект образует свой собственный кластер с единственным объектом. Таким образом, если известна матрица попарных расстояний, то наивная реализация алгоритм имеет сложность O(n^3).
В случае произвольной матрицы расстояний не существует решения работающего o(n^2), то есть быстрее линии от размера матрицы. Таким образом оптимизация этого алгоритма возможно только в частных случаях. Существует ряд работ, об ускоренном UPGMA для расстояния Минковского, расстояния Хэмминга, косинусной похожести. В большей части существующих работ не рассматривается фактор оптимизации метрик иерархической кластеризации как NP полной задачи.

Основная часть. Основная сложность UPGMA в поиске пары ближайших кластеров без перебора всех пар. Главный вопросом работы найти ограничения на матрицу расстояний/похожестей, при которых возможно выполнить все шаги алгоритма суммарно за O(n^(2-const)) и имеет теоретические гарантии аппроксимации широко используемой метрики иерархической кластеризации Mosgly-Wang objective.
Один из важных и ранее неизученных функций похожести являются ядра, нашедшие широкое применение в машинном обучение. Например, одно из наиболее популярных ядер похожести между векторами - RBF.

В отличие от классического подхода использования ядер, мы используем рандомизированный ядерный метод, который позволяет находить явное приближенное отображение векторное в пространство скалярный произведений ядра. В таком пространстве для вычисления ядра между векторами достаточно найти скалярное произведение. Для некоторых ядер существует точное конечномерное отображение, например: полиномиальное ядро, косинусное ядро, линейное ядро.
Явное пространство для ядра является линейным по определению. Это позволяет считать среднее значение ядра между всеми парами точек принадлежащих двум множествам. Предположим, что kernel(a, b) = <F(a), F(b)>, тогда:

.
.
. $$\frac{1}{|X| \cdot |Y|} \sum_{a \in X} \sum_{b \in Y} kernel(a, b) = \frac{1}{|X| \cdot |Y|} \sum_{a \in X} \sum_{b \in Y} \langle F(a), F(b) \rangle = \langle \frac{1}{|X|} \sum_{a \in X} F(a), \frac{1}{|Y|} \sum_{b \in Y} F(b) \rangle$$
.
.
Среднее значение точек после отображение является характеризующей точкой множества, которая задает его основные свойства. Таким образом для выполнения одного шага UPGMA необходимо найти пару точек с максимальным скалярным произведением. К сожалению, данная задача без дополнительных ограничений эквивалента проблеме ортогональных векторов, которая не может быть решена точно за O(n^(2-const)). Одно из возможных предположений это ограничение значения функции ядра снизу, таким образом мы ограничиваем снизу и значение скалярных произведений в пространстве отображений. Это условие позволяет решать задачу за O(n^(1-const)), используя структуру данных построенную на основе LSH и дискретизации длин векторов.

Дано: множество точек X, kernel с явным отображением F, mxdot — структура данных для поиска максимального скалярного произведения
Результат: Корень дерева иерархической кластеризации
Алгоритм:
```
for i from 0 to |X|
    mxdot.add(F(X[i]), Leaf(i))
for i from 0 to |X| - 1
   pair = mxdot.find_max_dot()
   mxdot.remove(pair.first)
   mxdot.remove(pair.second)
   new_node = Node(pair.fisrt.node, pair.second.node)
   f_sz = size(pair.fisrt.node)
   s_sz = size(pair.second.node)
   mxdot.add((pair.first.point * f_sz + pair.second.point * s_sz) / (f_sz + s_sz), new_node)
return mxdot.front().node
```

Можно доказать, что если структура данных находит ответ не более чем в 1-eps раз меньше, чем наибольшее скалярное произведение и минимальное значение kernel между всеми парами точек это alpha, то данный алгоритм имеет константный коэффициент аппроксимации Mosgly-Wang objective равный k = (1 + 2 * alpha) / (3 + eps), то есть результат не более чем в k раз хуже, чем оптимальный.

Выводы.
Предложен алгоритм, позволяющий получать пошаговое приближение результатов UPGMA для ядерных функций используя субквадратичное время и память.

Наумов С.С. (автор)

Фильченков А.А (научный руководитель)