И. С. С и м а к и н (Ярославль, ЯрГУ)**. Организация быстрого поиска ближайшего к заданной точке кластера на графе нестрогой иерархической кластеризации.**

Пусть задано множество точек-векторов в n-мерном пространстве. Кластеризация заключается в разбиении этого множества на кластеры таким образом, чтобы близкие точки содержались в одном кластере, а дальние — в различных [1]. Рассмотрим задачу, когда требуется выполнить разбиение на большое количество кластеров а затем обеспечить возможность быстрого поиска кластера, ближайшего к произвольно заданной точке, не выполняя перебор всех кластеров. Поставленную задачу можно эффективно решить при помощи алгоритма Local Sensitive Hashing [2]. В настоящей работе предложен альтернативный эффективный подход к решению этой задачи.

Итак, при выполнении иерархической кластеризации [1] строится дерево разбиения на кластеры, где каждый кластер верхнего уровня разбивается на кластеры нижнего уровня. В этом случае поиск ближайшего кластера сводится к движению по дереву сверху вниз, что требует логарифмического количества действий. Основным недостатком такого подхода является то, что решение, к какому подмножеству кластеров ближе всего заданная точка, принятое при движении по верхнему уровню дерева, нельзя пересмотреть на более низком уровне, учитывающем большее количество факторов.

Для преодоления данного недостатка воспользуемся обычным методом кластеризации, например, k-means [1], после чего построим *граф нестрогой иерархической кластеризации* (Nonstrict Hierarchical Clasterization Graph). На самом нижнем уровне иерархии в качестве вершин возьмем центры всех кластеров, а ребрами соединим соседние кластеры. В следующем уровне оставим некоторую часть вершин предыдущего уровня, образующую разреженную сеть, и т. д., пока не останется всего один корневой кластер верхнего уровня.

Поиск кластера, ближайшего к заданной точке, на таком графе, осуществляется по уровням, начиная с верхнего. На каждом уровне выполняется движение по ребрам графа от начальной вершины до вершины на этом уровне, ближайшей к заданной точке. Каждый шаг выполняется в сторону уменьшения расстояния до заданной точки. После этого осуществляется переход к более низкому уровню, и поиск продолжается.

Наличие ребра (отношение соседства) между вершинами определяется необходимостью этого ребра для обеспечения достижимости из корневой вершины ближайших кластеров для каждой точки из исходной выборки в процессе поиска.

Предложенную иерархическую кластеризацию следует называть *нестрогой*, потому что до одной вершины нижнего уровня можно добраться по нескольким путям через различные вершины более высокого уровня. При этом поиск имеет логарифмическую трудоемкость. При проведении компьютерного эксперимента на искусственных данных на один поиск по 1000 кластерам пришлось в среднем около 27 вычислений расстояния.

Предложенную нестрогую иерархическую кластеризацию можно применять для ускорения вычислений при анализе данных. Например, аппроксимация трудоемких методов классификации таких, как метод k-того соседа, позволила бы существенно увеличить их быстродействие.

СПИСОК ЛИТЕРАТУРЫ

1. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning: Data Mining, Inference and Prediction. Berlin etc.: Springer, 2009, 746 p.
2. *Andoni A.* Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions./ A. Andoni and P. Indyk. — Communications of the ACM, 51(1), 2008, p. 117–122.