Модели кластеризации

Айдагулов Р.Р., Главацкий С.Т., Михалев А.В.

Принято считать, что термин «кластеризация» (сгусток, пучок), был предложен математиком Р. Трионом. Впоследствии возник целый ряд терминов, которые рассматриваются как синонимы термина «кластерный анализ» или «автоматическая классификация». У кластерного анализа очень широкий спектр применения, его методы используются в медицине, химии, археологии, маркетинге, геологии и других дисциплинах. Кластеризация состоит в объединении в группы схожих объектов, и эта задача является одной из фундаментальных в области анализа данных. Обычно под кластеризацией понимается разбиение заданного множества точек некоторого метрического пространства на подмножества таким образом, чтобы близкие точки попали в одну группу, а дальние – в разные. Такое требование является достаточно противоречивым. Интуитивное разбиение «на глаз» использует соображение связности получаемых групп, исходя из плотности распределения точек. В докладе предлагается метод кластеризации, основанный на этой идее.

Определим плотность распределения заданного множества из N точек в точке x евклидова пространства методом локального осреднения следующим образом. Пусть точка x_i включена в шар единичного объема с центром в точке x с вероятностью $P(x, x_i)$. Тогда среднее количество (математическое ожидание) точек, входящих в этот объем, равно $\sum_i P(x, x_i)$. Если $\int P(x, y)\, dy = 1$, то полученное среднее и есть плотность (среднее количество точек в шаре единичного объема с центром в точке x). Для определения областей сгущения плотности нам требуется такое «локальное» осреднение. Далее мы можем так же вычислять и среднее количество точек, входящее в шар радиуса R (радиус осреднения) с центром в точке x.

Радиус осреднения для всей совокупности N точек в d–мерном пространстве следует выбирать так, чтобы, с одной стороны, среднее количество точек n в одном шаре такого радиуса было намного больше, чем $(\ln ln\, N)^d$, а, с другой стороны, – намного меньше, чем N^{ε}. Это свойство выполняется для такой, часто встречающейся функции, как $n = L(N) = exp\left(\sqrt{(\ln ln\, N)(\ln ln\ (\ln ln\, N))}\right)$.

Метод осреднения заключается в осреднении множества точек с функцией плотности распределения $\sum_i \delta(x - x_i)$. Выбирая далее срезы множества точек

по определенному уровню плотности, мы получим разбиение на кластеры. Этот метод свободен от таких недостатков, как зависимость от нумерации точек, и как существенное изменение разбиения на кластеры при малом изменении позиции даже одной точки.