Анализ массивов данных на базе простых подходов,
применимых в процедурах кластеризации

Алиев К.К.
«Липецкий государственный технический университет» Липецк, Россия

В последние годы, в связи с появлением *WWW* поисковых систем, и особенно проблемы организации огромного количества информации, и концепцией «информационная проходка» снова возник интерес к алгоритмам кластеризации. Напомним, что алгоритмы кластеризации бурно исследовались еще в *80*-х годах.

В большинстве случаев огромные объемы информации можно сделать доступными для восприятия, если уметь разбить источники информации на тематические группы [1]. Этот процесс осуществляется за счет процедуры кластеризации.

Задача кластеризации заключается в следующем. Имеется обучающая выборка $X_\ell = \{x_1,\ldots, x_\ell\} \in X$ и функция расстояния между объектами $\rho(x, x')$. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X_\ell$ приписывается метка (номер) кластера y_i.

Существует множество алгоритмов кластеризации, но большая их часть основана на представлении выборки в виде графа. Вершинам графа соответствуют объекты выборки, а рёбрам − попарные расстояния между объектами $\rho_{ij} = \rho(x_i, x_j)$. Достоинством графовых алгоритмов кластеризации является наглядность, относительная простота реализации, возможность вносить различные усовершенствования, опираясь на простые геометрические соображения.

Решается задача «двух связистов» [2]: два связиста играют в игру. Имеется n телефонных узлов, и связисты по очереди соединяют кабелями два из них по своему выбору. Выигрывает тот, после хода которого с любого узла можно будет дозвониться до любого другого (может быть, через несколько промежуточных). Узлами являются исходные кластеры.

Очутившись в рамках такой ситуации, нетрудно определить «победителя». При $n = 4,7,8$ выигрывает первый игрок, при $n=5,6$ – второй; это будет следовать из решения общей задачи.

Будем называть группу узлов связной, если любые два входящих в нее узла соединены (может быть через несколько других узлов). Связную группу узлов будем называть компонентой, если ни один из входящих в нее узлов не соединен ни с одним из узлов, в ней не содержащихся. Ясно, что любой граф (система как-то соединенных узлов) состоит из одной или нескольких компонент.

При любой игре противников настанет момент, когда множество всех узлов окажется разбитым на три компоненты A_1, A_2, A_3. Пусть a_1, a_2, a_3 – количества узлов в этих компонентах. Ясно, что проигрывает тот из игроков, кому придется первым соединить какие-то две из этих компонент. Критическая ситуация возникает, когда все возможные соединения внутри компонент A_1, A_2, A_3 будут осуществлены.

Ясно, что с начала партии до наступления такого момента будет сделано

$$N = \frac{a_1(a_1-1)}{2} + \frac{a_2(a_2-1)}{2} + \frac{a_3(a_3-1)}{2} = \frac{a_1^2 + a_2^2 + a_3^2 - n}{2}$$

ходов. Если число N окажется четным, то первый игрок проигрывает, если же нечетным – то выигрывает. Четность числа N, очевидно, определяется остатком, который дает число $a_1^2 + a_2^2 + a_3^2 - n$ при делении на 4. Так как остаток от деления на 4 числа $a_1^2 + a_2^2 + a_3^2$ равен количеству нечетных чисел среди чисел a_1, a_2, a_3, первый игрок должен добиться того, чтобы количество нечетных чисел среди a_1, a_2, a_3 не совпало с остатком от деления n на 4. Второй же, наоборот, должен стремиться к их совпадению.

В дальнейшем компоненты с четным числом узлов мы будем называть четными, а с нечетным числом узлов – нечетными.

Рассмотрим отдельно четыре случая: $n = 4k$, $n = 4k + 1$, $n = 4k + 2$, $n = 4k + 3$.

1. $n = 4k$. Первый игрок добивается победы, играя следующим образом. Пока компонент более трех, он играет так, чтобы после каждого его хода получилось не более одной четной компоненты: если в какой-то момент возникают две четные компоненты, он тут же объединяет одну из них с нечетной. В момент появления трех компонент (независимо от того, после чьего хода это произойдет) две компоненты будут нечетными, а одна четной. Число N при этом нечетно – первый игрок выигрывает.
2. $n = 4k + 1$. В этом случае выигрывает второй игрок, поскольку он может играть так, чтобы после каждого его хода было не менее двух четных компонент. При этом среди чисел a_1, a_2, a_3 два непременно будут четными, а одно – нечетным (независимо от того, чей ход приведет к образованию трех компонент).
3. $n = 4k + 2$. И в этом случае выигрывает второй игрок, применяя стратегию первого игрока в случае $n = 4k$, так как при $n = 4k + 2$ среди чисел a_1, a_2, a_3 будут два нечетных.
4. $n = 4k + 3$. Этот случай наиболее интересен.

   При $n = 3$, очевидно, выигрывает второй игрок. Оказывается, что при $n > 3$ выигрывает второй игрок.

Фактически это означает то же самое, что и пусть рассмотрим полный граф с N вершинами. Каждой вершине соответствует объект. Припишем ребрам графа веса d_{ij}.
Пусть r — некоторый параметр. Удалим из графа все ребра, веса которых больше r.

Компоненты связности полученного таким образом графа и есть кластеры. Для полного графа из N вершин, ребрам которого приписаны веса d_{ij}, построим минимальное остовное дерево. Удалим из дерева K ребер максимального веса. Получим разбиение множества объектов на K кластеров.

Рассмотрение таких подобных «простых», соблазнительных задач, как эта, дает нам уверенность в преодолении трудностей. Таких задач, на которые можно опереться, много. Это задача Колмогорова [3]: фиксируем на плоскости угол и точку внутри него. Ставится задача существования фигуры, отличного от круга, способной вращаться внутри данного угла так, чтобы граница фигуры все время касалась сторон угла и проходила через точку; задача «Царевны Дидоны» - задача изопериметров и мыльных пленок [4]; задача «медианная фильтрация», которая рассматривает грубые ошибки данных (выбросы), среднее значение и медиану, обработку дискретного сигнала, инвариантные последовательности и обработку изображений.

Таким образом, решение задачи кластеризации, в общем виде, принципиально неоднозначно, и тому есть несколько причин. Во-первых, не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд достаточно разумных критериев, а также ряд алгоритмов, не имеющих чётко выраженного критерия, но осуществляющих достаточно разумную кластеризацию по построению [5]. Все они могут давать разные результаты. Во-вторых, число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием. В-третьих, результат кластеризации существенно зависит от метрики ρ, выбор которой, как правило, также субъективен и определяется экспертом.

Список использованных источников

1. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.
2. Разборов А. Задача о двух связистах// Научно-популярный физико-математический журнал «КВАНТ». -1981.- №12. –С. 29-30.
3. Бабичев А. Об одной задаче Колмогорова // Научно-популярный физико-математический журнал «КВАНТ». -1981.- №5. –С. 14-16.
4. Трофимов В.В. Царевна Дидона, изопериметры и мыльные пленки // Научно-популярный физико-математический журнал «КВАНТ». -1985.- №5. –С. 22-27.
5. Айвазян А.С., Мхитарян В.С. Прикладная статистика и основы эконометрики. М.: Издательское объединение ЮНИТИ, 1998.