УДК 004.9

МЕТОДЫ КЛАСТЕРИЗАЦИИ МНОГОМЕРНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ

Князь Д.В.,
научный руководитель кандидат физ-мат. наук, доцент Баранова И. В.
Сибирский Федеральный Университет
Институт математики и фундаментальной информатики

Введение

Кластерный анализ (*cluster analysis*) — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя, термин впервые ввел Трион в 1939 году.

Под *кластером* обычно понимается часть данных (в типичном случае – подмножество объектов или подмножество переменных, или подмножество объектов, характеризуемых подмножеством переменных), которая выделяется из остальной части наличием некоторой однородности ее элементов. В простейшем случае речь идет о похожести элементов, в идеальном случае – о совпадающих значениях основных переменных или иного рода близости, выражаемой геометрической близостью соответствующих объектов.

Формальная постановка задачи кластеризации: Пусть X — множество объектов, Y — множество номеров (имён, меток) кластеров. Задана функция расстояния между объектами $p(x, x')$. Имеется конечная обучающая выборка объектов $X^m = x_1, \dots, x_m \subset X$. Требуется разбить выборку на непересекающиеся подмножества, называемые *кластерами*, так, чтобы каждый кластер состоял из объектов, близких по метрике p, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера y_i.

Алгоритм кластеризации — это функция $a: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$. Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного *критерия качества* кластеризации.

Кластеризация (обучение без учителя) отличается от классификации (обучение с учителем) тем, что метки исходных объектов y_i изначально не заданы, и даже может быть неизвестно само множество Y.

Методы кластеризации

Общепринятой классификации методов кластеризации не существует, но можно выделить ряд общепринятых групп подходов:

1. Вероятностный подход: K-средних (K-means), K-medians, EM-алгоритм, Алгоритмы семейства FOREL, Дискриминантный анализ
2. Подходы на основе систем искусственного интеллекта: Метод нечеткой кластеризации C-средних (C-means), Нейронная сеть Кохонена, Генетический алгоритм
3. Логический подход: деревья решений и т.п.
4. Теоретико-графовый подход:
    a. Графовые алгоритмы кластеризации
5. Иерархический подход

6. Другие методы:
    a. Статистические алгоритмы кластеризации
    b. Ансамбль кластеризаторов
    c. Алгоритмы семейства KRAB
    d. Алгоритм, основанный на методе просеивания.

Выбор метода кластеризации зависит от количества данных и от того, есть ли необходимость работать одновременно с несколькими типами данных.

Теоретико-графовый и иерархический подходы иногда объединяют под названием структурного или геометрического подхода, обладающего большей формализованностью понятия близости. Несмотря на значительные различия между перечисленными методами все они опираются на исходную «**гипотезу компактности**»: в пространстве объектов все близкие объекты должны относиться к одному кластеру, а все различные объекты соответственно должны находиться в различных кластерах.

В основе теоретико-графовых методов лежит принцип представления данных в виде вершин графа, а расстояние между ними имеет смысл ребер графа. Для последующей кластеризации используется сравнительно широкий спектр методов, разработанный для анализа графов. Типичными представителями графовых алгоритмов кластеризации являются: алгоритм кратчайшего незамкнутого пути, выделения связных компонент, ФОРЭЛ и другие.

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделение больших кластеров на меньшие.

К иерархическим методам относятся метод ближайшего соседа (или одиночная связь), метод наиболее удаленных соседей (или полная связь), метод Варда, метод невзвешенного попарного среднего, метод взвешенного попарного среднего, невзвешенный центроидный метод,взвешенный центроидный метод и другие.

Сложности иерархических методов кластеризации: ограничение объема набора данных; выбор меры близости; негибкость полученных классификаций. Преимущество этой группы методов – их наглядность и возможность получить детальное представление о структуре данных.

При большом количестве наблюдений иерархические методы кластерного анализа не пригодны. В таких случаях используют неиерархические методы, основанные на разделении, которые представляю собой итеративные методы дробления исходной совокупности. В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки. Неиерархические методы выявляют более высокую устойчивость по отношению к шумам и выбросам, некорректному выбору метрики, включению незначимых переменных в набор, участвующий в кластеризации.

В работе рассматриваются три наиболее популярных алгоритма кластеризации данных: метод К-средних, кратчайшего незамкнутого пути и Ланса-Уильямса. Приведем описания перечисленных алгоритмов кластеризации.

Метод К-средних имеет следующий вид:

Предположим, что мы имеем N объектов имеющих n переменных, которые должны быть разделены на К кластеров.

1.Выбирается число Z, и на первом шаге эти точки считаются "центрами" кластеров. Каждому кластеру соответствует один центр. Выбор начального центроида осуществляется случайным образом.

2. Вычисляется минимальное p_{ij} (p_{ij} -расстояние Евклида, может быть другое) от переменных до центроидов. В результате каждый объект назначен определенному кластеру.

3. Вычисляются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются. Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий: а) кластерные центры стабилизировались, т.е все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации; б) число итераций равно максимальному числу итераций.

Выбор числа кластеров является сложным вопросом. Если нет предположений относительно этого числа, рекомендуется создать 2 кластера, затем 3,4,5 и т.д., сравнивая полученные результаты.

Алгоритм КНП - "Кратчайший Незамкнутый Путь":

1. Найти пару вершин с наименьшим p_{ij}, где p_{ij} - расстояние между вершинами, и соединить их ребром;
2. Пока в выборке остаются изолированные точки
3. Найти изолированную точку, ближайшую к некоторой неизолированной;
4. Соединить эти две точки ребром;
5. Удалить К-1 самых длинных ребер, К-число кластеров.

Алгоритм Ланса-Уильямса:

1. Сначала все кластеры одноэлементные: $t = 1,\ R(\{x_i\},\{x_j\}) = \rho(x_i, x_j)$.
2. Для всех t=2,...,l (t - номер итерации);
   2.1 Найти в C_{t-1} два ближайших кластера: $(U,V) := \arg\ min_{U \neq V}$ R(U,V) ; $R_t := R(U,V)$
   2.2. Слить их в один кластер: $W := U \cup V$; $C_t := C_t(t-1) \cup \{W\}\{U,V\}$;
   2.3. Для всех $S \in C_t$ вычислить R(W,S) по формуле Ланса-Уильямса

Формула Ланса-Уильямса:

$$R(U \cup V, S) = \alpha_U \cdot R(U,S) + \alpha_V \cdot R(V,S) + \beta \cdot R(U,V) + \gamma \cdot |R(U,S) - R(V,S)|,$$

где $\alpha_U, \alpha_V, \beta, \gamma$ – числовые параметры.

Частные случаи формулы Ланса-Уильямса:

а) Расстояние ближнего соседа: $R^{б}(W,S) = min_{w \in W, s \in S}\ p(w,s)$,

$$\alpha_U = \alpha_V = \frac{1}{2}, \qquad \beta = 0, \qquad \gamma = -\frac{1}{2}$$

б) Расстояние дальнего соседа: $R^{д}(W,S) = max_{w \in W, s \in S}\ p(w,s)$,

$$\alpha_U = \alpha_V = \frac{1}{2}, \qquad \beta = 0, \qquad \gamma = \frac{1}{2}$$

в) Групповое среднее расстояние: $$R^{г}(W,S) = \frac{1}{|W||S|}\sum_{w \in W}\sum_{s \in S} p(w,s)$$;

$$\alpha_U = \frac{|U|}{|W|}, \qquad \alpha_V = \frac{|V|}{|W|}, \qquad \beta = \gamma = 0$$

г) Расстояние между центрами: $R^{ц}(W,S) = p^2(\sum_i(w \in W)w/|W| \cdot \sum_i(s \in S)s/(|S|)$);

$$\alpha_U = \frac{|U|}{|W|}, \qquad \alpha_V = \frac{|V|}{|W|}, \qquad \beta = -\alpha_U \alpha_V, \qquad \gamma = 0$$

д) Расстояние Уорда: $$R^{у}(W,S) = \frac{|S||W|}{|S|+|W|}p^2\left(\sum_{w \in W}\frac{w}{|W|}, \sum_{s \in S}\frac{s}{|S|}\right)$$;

$$\alpha_U = \frac{|S|+|U|}{|S|+|W|}, \qquad \alpha_V = \frac{|S|+|V|}{|S|+|W|}, \qquad \beta = \frac{-|S|}{|S|+|W|}, \qquad \gamma = 0$$

В работе было проведено сравнение методов кластеризации между собой, с точки зрения требований к входным данным и получаемых кластеров (в т..ч. формы

кластеров и точность полученного разбиения данных). Полученные результаты приведены в таблице 1.

Таблица 1

Сравнительная таблица алгоритмов кластеризации

Алгоритм кластеризации	**Форма кластеров**	**Входные данные**	**Результаты**
Иерархический	Произвольная	Число кластеров или порог расстояния для усечения иерархии	Бинарное дерево кластеров
k-средних	Гиперсфера	Число кластеров	Центры кластеров
c-средних	Гиперсфера	Число кластеров, степень нечеткости	Центры кластеров, матрица принадлежности
Выделение связных компонент	Произвольная	Порог расстояния R	Древовидная структура кластеров
Минимальное покрывающее дерево	Произвольная	Число кластеров или порог расстояния для удаления ребер	Древовидная структура кластеров
Послойная кластеризация	Произвольная	Последовательность порогов расстояния	Древовидная структура кластеров с разными уровнями иерархии

Практическая задача кластеризации

В работе решается практическая задача кластеризации ведущих российских банков по основным показателям их финансовой стабильности. Для решения задачи использовалась официальная статистика отчетности кредитных организаций РФ по показателям их деятельности, публикуемая на сайте Банка России. В данной работе решается задача кластеризации десяти российских банков. В статистике рассматриваются следующие основные показатели деятельности банков: активы нетто, чистая прибыль, кредитный портфель, просроченная задолженность в кредитном портфеле, вклады физических лиц, вложение в ценные бумаги, средства предприятий и организаций, привлеченные межбанковские кредиты, векселя и облигации.

Кластеризация данных проводилась с помощью трёх методов: метода К-средних (вероятностный подход); с использованием алгоритма КНП - "Кратчайший незамкнутый путь" (теоретико-графовый подход), и алгоритма Ланса-Уильямса (иерархический подход). Для алгоритма Ланса-Уильямса рассматривались частные случаи формулы Ланса-Уильямса, такие как:

а) расстояние ближнего соседа;

б) групповое среднее расстояние;

в) расстояние между центрами.

Также в работе было проведено сравнение кластеров, полученных в результате работы каждого метода.