Д.Ю. КУЗНЕЦОВ, Т.Л. ТРОШИНА

Кластерный анализ и его применение

Исследователь часто стоит перед лицом огромной массы индивидуальных наблюдений. Возникает задача сведения множества характеристик к небольшому ряду обобщающих итогов, выражающему действительно существенное для явления. Но пока каждый вовлеченный в анализ признак остается отдельным самостоятельным элементом со своими характеристиками, число параметров, выражающих результаты обработки, не поддается уменьшению. Единственный путь к нему – либо в отсечении большинства признаков и возвращении к малоразмерным классическим задачам, либо в объединении признаков, в замене целых «гроздей» их одним, искусственно построенным на их основе. Так и появилось направление – «многомерный анализ».

В многомерном статистическом анализе образовались разделы, которые не изолированы, а проникают, переходят один в другой. Это кластерный анализ, метод главных компонент, факторный анализ. Наиболее ярко отражают черты многомерного анализа в классификации объектов кластерный анализ, а в исследовании связей – факторный анализ.

Кластерный анализ – это способ группировки многомерных объектов, основанный на представлении результатов отдельных наблюдений точками подходящего геометрического пространства с последующим выделением групп как «сгустков» этих точек (кластеров, таксонов). «Кластер» (cluster) в английском языке означает «сгусток», «гроздь винограда», «скопление звезд» и т.д. Данный метод исследования получил развитие в последние годы в связи с возможностью компьютерной обработки больших баз данных.

Кластерный анализ предполагает выделение компактных, удаленных друг от друга групп объектов, отыскивает «естественное» разбиение совокупности на области скопления объектов. Он используется, когда исходные данные представлены в виде матриц близости или расстояний между объектами либо в виде точек в многомерном пространстве. Наиболее распространены данные второго вида, для которых кластерный анализ ориентирован на выделение некоторых геометрически удаленных групп, внутри которых объекты близки.

Выбор расстояния между объектами является узловым моментом исследования, от него во многом зависит окончательный вариант разбиения объектов на классы при данном алгоритме разбиения.

Существует большое количество алгоритмов кластерного анализа, их можно разделить по способу построения кластеров на 2 типа: эталонные и неэталонные. В процедурах эталонного типа на множестве объектов задается несколько исходных зон, с которых начинает работу алгоритм. Эталоны могут представлять собой первоначальное разбиение на классы, центр

тяжести класса и др. После задания эталонов алгоритм производит классификацию, иногда меняя определенным способом эталоны.

К алгоритмам кластеризации, работающим по иному принципу, относятся иерархические алгоритмы кластерного анализа, процедура разрезания и др.

Задача кластерного анализа

Пусть множество $I=\{I_1,I_2,\ldots,I_n\}$ обозначает n объектов. Результат измерения i-й характеристики I_j объекта обозначают символом x_{ij}, а вектор $X_j=[x_{ij}]$ отвечает каждому ряду измерений (для j-го объекта). Таким образом, для множества I объектов исследователь располагает множеством векторов измерений $X=\{X_1, X_2,\ldots,X_n\}$, которые описывают множество I. Множество X может быть представлено как n точек в p-мерном евклидовом пространстве E_p.

Пусть m – целое число, меньшее чем n. Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве X, разбить множество объектов I на m кластеров (подмножеств) $\pi_1,\pi_2,\ldots, \pi_m$ так, чтобы каждый объект I_j принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие разным кластерам, были разнородными (несходными).

Решением задачи кластерного анализа является разбиение, удовлетворяющее некоторому условию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и группировок. Этот функционал часто называют целевой функцией. Задачей кластерного анализа является задача оптимизации, т.е. нахождение минимума целевой функции при некотором заданном наборе ограничений. Примером целевой функции может служить, в частности, сумма квадратов внутригрупповых отклонений по всем кластерам.

Основные понятия кластерного анализа

N измерений $X_1, X_2,\ldots,X_n$ могут быть представлены в виде матрицы

$$X=[X_1, X_2,\ldots,X_n]=\begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1N} \\ x_{21} & x_{22} & \ldots & x_{2N} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nN} \end{bmatrix}.$$

Аналогичным образом расстояния между парами векторов $d(X_i,X_j)$ могут быть представлены в виде матрицы расстояний:

$$\Delta=\begin{bmatrix} 0 & d_{12} & \ldots & d_{1n} \\ d_{21} & 0 & \ldots & d_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ d_{n1} & d_{n2} & \ldots & 0 \end{bmatrix},$$

$d_{ii}=0$ для i=1,2,…,n.

Если признаки измерены в разных единицах измерения, то определить расстояние между объектами нельзя. Тогда применяется нормировка показателей, переводящая их в безразмерные величины. Наиболее распространенные способы нормирования следующие:

$$z1 = \frac{x - \bar{x}}{\sigma},\ z2 = \frac{x}{\bar{x}},\ z3 = \frac{x}{x_{\max}},\ z4 = \frac{x - \bar{x}}{x_{\max} - x_{\min}}.$$

Понятием, противоположным понятию расстояния между объектами X_i и X_j, является понятие близости (сходства) между X_i и X_j. Точнее, мера близости между объектами X_i и X_j – это вещественная функция $\mu(X_i,X_j)=\mu_{ij}$ со свойствами:

$$0 \le \mu(X_i,X_j) < 1 \text{ для } X_i \ne X_j;$$
$$\mu(X_i,X_i)=1;$$
$$\mu(X_i,X_j)=\mu(X_j,X_i).$$

Пары значений мер близости можно объединить в матрицу близости:

$$\mu=\begin{bmatrix} 1 & \mu_{12} & \dots & \mu_{1n} \\ \mu_{21} & 1 & \dots & \mu_{2n} \\ \dots & \dots & \dots & \dots \\ \mu_{n1} & \mu_{n2} & \dots & 1 \end{bmatrix},\ \mu_{ii}=0 \text{ для } i=1,2,\dots,n.$$

Величину μ_{ij} называют коэффициентом близости. Примером линейной близости является коэффициент корреляции.

Рассмотрим основные способы определения расстояний между объектами.

Метрики для количественных шкал (расстояние).

а) Линейное расстояние

$$d(X_j,X_i)=\sum_{k=1}^{N} | x_{ki} - x_{kj} |;$$

б) евклидово расстояние

$$d(X_j,X_i)=\left[\sum_{k=1}^{N} (x_{ki} - x_{kj})^2\right]^{1/2};$$

в) обобщенное степенное расстояние Минковского (универсальная метрика)

$$d(X_j,X_i)=\left[\sum_{k=1}^{N} (x_{ki} - x_{kj})^p\right]^{1/p}.$$

Метрики для качественных шкал (мера близости).

К качественным шкалам относят:

а) номинальную шкалу (или шкалу наименований). Примеры измерения: пол (мужчина, женщина), национальность (француз, итальянец, немец), профессия (учитель, врач, бухгалтер) и др.;

б) порядковую шкалу (или ранговую, ординарную). Примеры измерения: экспертные ранжировки, оценки предпочтений, шкала твердости минералов и др.

Расстояние для номинальных шкал вводится следующим образом. Пусть имеются два объекта X и Y с N признаками. Введем координаты x_i и y_i (i=1,2,…,N) как логические переменные, принимающие значение 1, если объект обладает i-м признаком, и 0, если признак с номером i у объекта отсутствует.

Выбор конкретного измерителя близости объектов X и Y должен осуществляться из содержательных соображений: если предполагается значимость совпадения единичных и нулевых свойств, то применяют расстояние Хемминга – отношение количества совпадающих значений к числу всех значений N. Если же важно наличие свойства, а не его отсутствие, то применяют коэффициенты Рао или Роджерса-Танимото, в которых учитываются только совпадающие единичные значения, а совпадающие нулевые игнорируются.

Матрицы расстояний Δ или близостей μ нередко задаются непосредственно либо как таблицы экспертных оценок близости, либо как матрицы прямых измерений сходства, например, матрицы межотраслевого баланса, степеней соседства географических регионов, взаимной цитируемости авторов и т.д.

Рассмотрим возможные способы точного определения кластеров.

Класс типа сгущения (класс типа ядра): все расстояния между объектами внутри класса меньше любого расстояния между объектами класса и остальной частью множества.

Класс с центром: класс называется классом с центром, если существует порог $\tau>0$ и некоторая точка x_l^* в пространстве, занимаемом объектами кластера S_l со свойствами:

если $d_{ix_l^*} \le \tau$, то $x_i \in S_l$;

если $d_{ix_l^*} > \tau$, то $x_i \notin S_l$.

Точка x_l^* называется центром класса. Часто в качестве x_l^* рассматривается центр тяжести, то есть координаты центра определяются как средние значения признаков у объектов класса.

Далее пусть $X=\{X_1, X_2,\ldots,X_{n_1}\}$ – множество измерений, произведенных над множеством объектов $I=\{I_1, I_2,\ldots,I_{n_1}\}$, а $Y=\{Y_1, Y_2,\ldots,Y_{n_2}\}$ – множество измерений, соответствующее множеству $J=\{J_1, J_2,\ldots,J_{n_2}\}$.

Величину $D=\frac{n_1 n_2}{n_1+n_2}(\overline{X}-\overline{Y})^T(\overline{X}-\overline{Y})$, где $\overline{X}=\sum_{i=1}^{n_1}\frac{X_i}{n_1}$, $\overline{Y}=\sum_{i=1}^{n_2}\frac{Y_i}{n_2}$ называют расстоянием между кластерами I и J. Формула вычисления расстояния между кластерами используется как один из параметров в алгоритмах кластерного анализа.

В настоящее время процедуры эталонного типа применимы для решения многих задач классификации, алгоритмы быстры и удобны в вычислительном отношении, их результаты наглядно представимы в диаграммах и графиках. Для проведения эталонной классификации

необходимо выбрать метод первичного задания эталонных множеств и способ корректировки классов и стабилизации в целом, задать значения параметров алгоритма кластеризации.

Иерархические алгоритмы кластерного анализа могут быть двух типов – агломеративные и дивизионные. В агломеративных процедурах начальным является разбиение, состоящее из n одноэлементных классов, а конечным – из одного класса, в дивизионных – наоборот. Принцип работы иерархических агломеративных (дивизионных) алгоритмов состоит в последовательном объединении (разделении) групп элементов, т.е. в создании иерархической структуры классов. Обычно такая классификация представляется в виде дендограммы – графика, отражающего последовательное объединение двух кластеров в один с указанием расстояний между ними.

В качестве частного примера рассмотрим результаты кластерного анализа, проведенного с использованием статистического пакета “Statistica”. Анализировались результаты тестирования (тест Амтхауэра на определение интеллектуального уровня, состоящий из 9 субтестов) и экспертная оценка успеваемости студентов ЯГПУ. Данные предварительно нормировались. При анализе определялся метод анализа, вид формулы для расстояния (евклидово) и количество кластеров (3) в эталонном алгоритме. Средние значения субтестов и экспертной оценки успеваемости (EXPOSR) для каждого кластера представлены на рис.1 (все курсы) и рис.2 (3 курс).

Их анализ наглядно демонстрирует, например, что если в целом студенты с низкой экспертной оценкой успеваемости имеют и более низкие показатели IQ, то высокие показатели IQ не столь однозначно связаны с успеваемостью – ситуация меняется от курса к курсу, и, например, на 3 курсе более успешной оказывается группа со средними значениями IQ. В качестве примера выполнения иерархического агломеративного алгоритма приведем дендограмму тех же показателей, отражающую структуру связей между субтестами и успеваемостью на исследуемой выборке (рис.3).

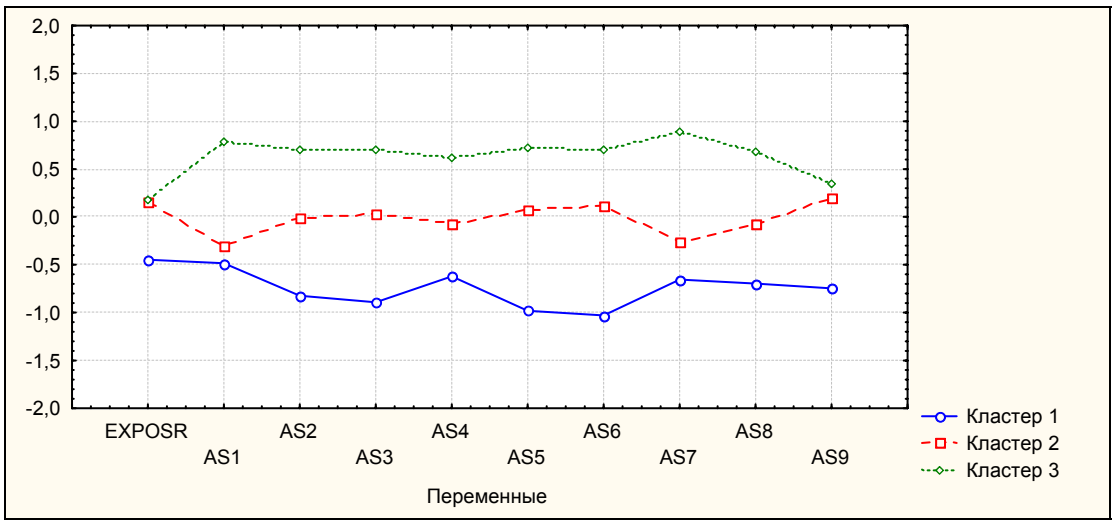

Рис.1

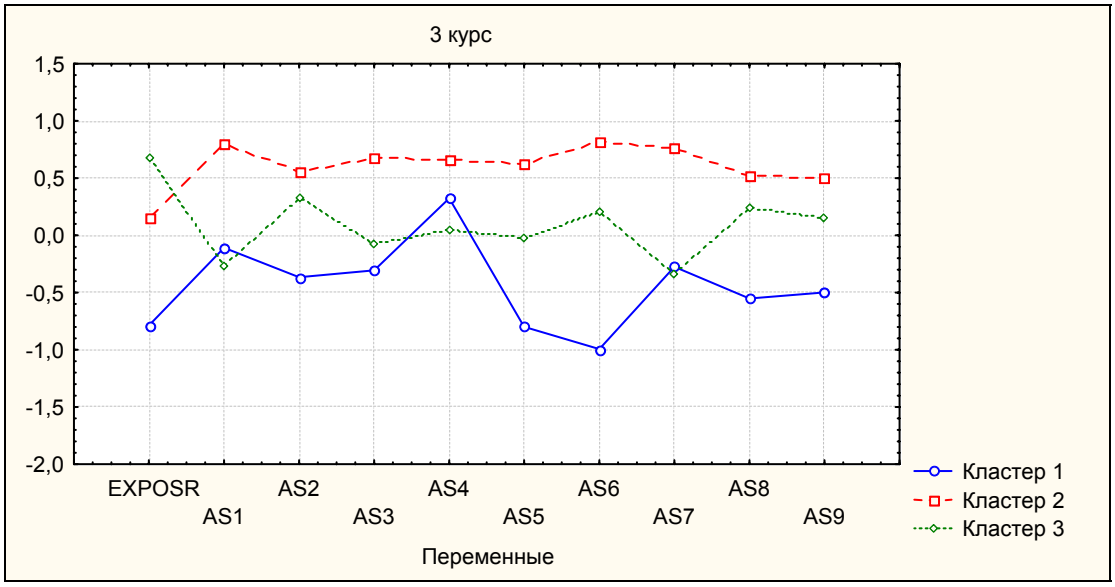

Рис.2

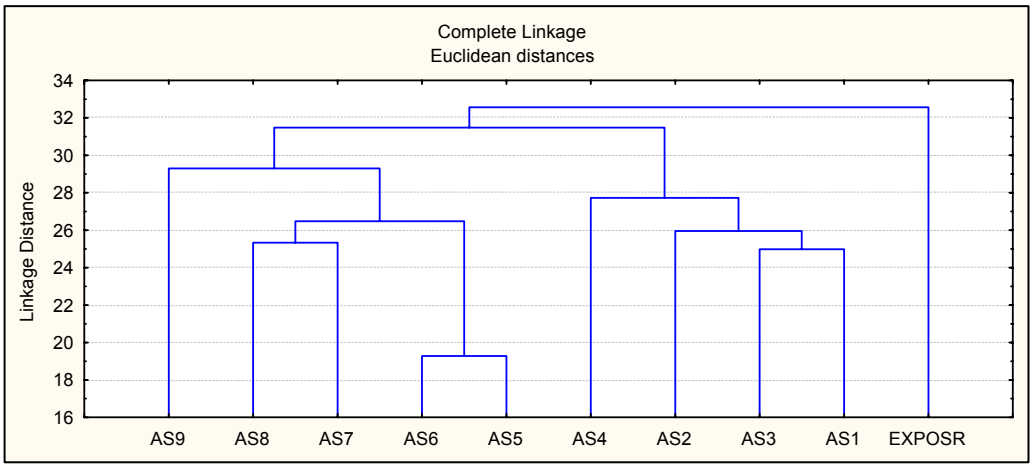

Рис.3

Библиографический список

1. Дюран Б., Оделл П. Кластерный анализ. М.: Статистика, 1977. 128 с.
2. Жамбю М. Иерархический кластер-анализ и соответствия. М.: Финансы и статистика, 1988. 342 с.
3. Мандель И.Д. Кластерный анализ. М.: Финансы и статистика, 1988. 176 с.