МАТЕМАТИЧЕСКИЕ ОСНОВЫ ТЕОРИИ ПРОГНОЗИРОВАНИЯ

Лектор
Сенько Олег Валентинович

Лекция 11

Методы кластерного анализа

Важное прикладное значение имеют методы анализа данных, связанные с теорией распознавания. К их числу относятся методы кластерного анализа и методы визуализации многомерных данных. Целью методов кластерного анализа является разбиение выборок многомерных данных на группы объектов близких в смысле некоторой заданной меры сходства. Такие компактные группы называются кластерами, классами или таксонами.

Методы кластерного анализа называют также методами обучения без учителя, автоматической группировки или таксономии

Методы кластерного анализа

Методы кластерного анализа могут использоваться в качества вспомогательных инструментов при решении задач прогнозирования или распознавания. Так с помощью кластеризации могут отбираться эталонные объекты. Однако нередко кластеризация может иметь самостоятельное значение.

Можно выделить задачи кластерного анализа, для которых число кластеров задано, а также задачи, в которых число кластеров следует определить в ходе решения кластеризации.

Методы кластерного анализа

Одним из наиболее известных методов кластеризации является **алгоритм *k внутригрупповых средних.*** Предположим, что у нас задана выборка многомерных векторов-объектов $\tilde{S}_{ini} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. Алгоритм находит такие кластеры, для объектов которых центр «своего кластера» будет ближе центра любого «чужого кластера».На начальном этапе произвольным образом выбирается начальная кластеризация $\tilde{G}_0 = \{G_1^0, \ldots, G_k^0\}$ с содержанием объектов $(m_1^0, \ldots, m_k^0)$ соответственно.

Методы кластерного анализа

Предположим, что на (*l-1*)-ом шаге получены группы $\tilde{G}_l = \{G_1^l, \ldots, G_k^l\}$

На *l*-ом Для каждой из групп G_i^l вычисляется центр

$$\bar{\mathbf{x}}_i^l = \frac{1}{m_i^l} \sum_{\mathbf{x}_j \in G_i^l} \mathbf{x}_j, \; i = 1, \ldots, k$$

Пусть $\rho(\mathbf{x}, \mathbf{y})$ – неотрицательная функция близости между векторами $\mathbf{x}, \mathbf{y}$

Произвольный объект $\mathbf{x}_{j'}$ переносится в группу $G_{i''}^l$, если

$$\rho(\mathbf{x}_{j'}, \bar{\mathbf{x}}_{i''}^l) < \rho(\mathbf{x}_{j'}, \bar{\mathbf{x}}_{i'}^l), \; i' \in \{1, \ldots, k\} \setminus i'' \; .$$

Методы кластерного анализа

В результате мы получаем группы $\tilde{G}_{l+1} = \{G_1^{l+1}, \ldots, G_k^{l+1}\}$ и переходим к (*l+1*)-ому шагу . Процесс останавливается, если на каком-то шаге оказывается, что $\overline{\mathbf{x}}_i^l = \overline{\mathbf{x}}_i^{l+1}, i = 1, \ldots, k$

Другим методом кластеризации, основанным на итерационной процедуре является алгоритм Форель, основанный на. движении гипершаров фиксированного радиуса в сторону мест «сгущения»объектов. Пусть фиксировано некоторое положительное число R. *Выбирается случайный вектор* $\mathbf{x}_{j^*} \in \tilde{S}_{ini}$ ***и*** гипершар радиуса R с центром в $\mathbf{z}_1 = \mathbf{x}_{j^*}$: $\mathbf{R}_1 = \{\mathbf{x} : \rho(\mathbf{x}, \mathbf{z}_1) < R\}$

Методы кластерного анализа

Полагаем $G_1 = \tilde{S}_{ini} \cap \mathbf{R}_1$ Вычисляется центр новой сферы

$\mathbf{z}_2 = \frac{1}{|G_1|} \sum_{\mathbf{x}_j \in G_1} \mathbf{x}_j$ и группа $G_2 = \tilde{S}_{ini} \cap \mathbf{R}_2$, где $\mathbf{R}_2 = \{\mathbf{x} : \rho(\mathbf{x}, \mathbf{z}_2) < R\}$

Процесс заканчивается на некотором шаге l^* при выполнении условия $G_{l^*+1} = G_{l^*}$. Полученное множество объектов объявляется первым кластером G_1^f. Оно исключается из $\tilde{S}_{ini}$ вышеописанная процедура повторяется относительно оставшейся части выборки.

Методы кластерного анализа

Таким образом последовательно находятся кластеры

$$G_1^f, G_2^f, \ldots, G_{k^*}^f$$

Процесс кластеризации заканчивается на k^* -ой итерации

при достижении условия $$\tilde{S}_{ini} \setminus \bigcup_{i=1}^{k^*} G_i^f = \varnothing$$

- Полученное число кластеров зависит от выбора радиуса *R, который является параметром* алгоритма..

Методы кластерного анализа

Метод иерархической группировки позволяет не только осуществить кластеризацию с заранее выбранным числом классов и выявить иерархию кластеров.

На начальном этапе в качестве кластеров рассматриваются отдельные объекты выборки $\tilde{S}_{ini}$. Дальнейшая кластеризация производится с используется функции близости между кластерами, которая задаётся на основе функции близости между векторными описаниями объектов.

Методы кластерного анализа

На практике используется несколько типов функций близости между кластерами $G_{i'}$ и $G_{i''}$:

$$P_{\min}(G_{i'}, G_{i''}) = \min_{\mathbf{x}_{\mu \in G_{i'}}, \mathbf{x}_{\nu \in G_{i''}}} \rho(\mathbf{x}_\mu, \mathbf{x}_\nu)$$ - минимальное расстояние между объектами из двух кластеров;

$$P_{\max}(G_{i'}, G_{i''}) = \max_{\mathbf{x}_{\mu \in G_{i'}}, \mathbf{x}_{\nu \in G_{i''}}} \rho(\mathbf{x}_\mu, \mathbf{x}_\nu)$$ - максимальное расстояние между объектами из двух кластеров;

$$P_c(G_{i'}, G_{i''}) = \rho(\bar{\mathbf{x}}_{i'}, \bar{\mathbf{x}}_{i''})$$ - расстояние между центрами двух кластеров;

Методы кластерного анализа

$$P_{av}(G_{i'}, G_{i''}) = \frac{1}{|G_{i'}| * |G_{i''}|} \sum_{\mu=1}^{|G_{i'}|} \sum_{\nu=1}^{|G_{i''}|} \rho(\mathbf{x}_\mu, \mathbf{x}_\nu)$$

- среднее

расстояние между объектами двух классов.

На втором шаге два ближайших кластера объединяются в один.

Процесс объединения повторяется до нахождения *l кластеров*.

Для остановки процесса объединения кластеров могут быть использованы дополнительные условия, задаваемые экспертом, и связанные со спецификой конкретной задачи.

В этом случае число кластеров устанавливается в ходе решения.

Методы кластерного анализа

Используются также методы кластеризации, основанные на поиске разбиений $\tilde{S}_{ini}$, для которых достигают максимума специальные функционалы качества. Так качество разбиения $\tilde{G} = \{G_1, \ldots, G_k\}$ может быть описано с помощью функционала внутренних дисперсий $F_{VS}(\tilde{G})$ представляющего собой взвешенную сумму средних отклонений от центра внутри внутри каждой из групп

$$F_{VS}(\tilde{G}) = \sum_{i=1}^{k} \{ |G_i| [\sum_{\mathbf{x}_j \in G_i}^{|G_i|} \frac{\rho(\mathbf{x}_j, \bar{\mathbf{x}}_i)}{|G_i|}] \} = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in G_i}^{|G_i|} \rho(\mathbf{x}_j, \bar{\mathbf{x}}_i)$$

Методы кластерного анализа

Нетрудно видеть, что “вес” каждой из групп пропорционален числу объектов в ней.

Поскольку число всевозможных разбиений $\tilde{S}_{ini}$ на k ***групп оценивается*** как $\frac{k^m}{k!}$ полный перебор разбиений здесь заведомо исключен. Поэтому обычно применяют методы частичного перебора с использованием случайного выбора начальных разбиений и последующей локальной оптимизацией

Методы кластерного анализа

В методах локальной оптимизации (для определенности, минимизации) строится последовательность разбиений

- $\tilde{G}_1, \tilde{G}_2, \ldots, \tilde{G}_l, \ldots$, для которых

$$F(\tilde{G}_1) > F(\tilde{G}_2), \ldots, F(\tilde{G}_l) > F(\tilde{G}_{l+1}), \ldots$$

а разбиение $\tilde{G}_{l+1}$ *вычисляется* непосредственно по $\tilde{G}_l = \{G_1^l, \ldots, G_k^l\}$ *путем его «локального» изменения – переноса некоторого объектов из одного кластера в другой.*

Методы кластерного анализа

Ищется такой объект $\mathbf{x}_{j_*(l)}$ *, при переносе которого из кластера* G_μ^l *(содержащего* $\mathbf{x}_{j_*(l)}$ *в разбиении* $\tilde{G}_l$ *) в некоторый кластер* G_ν^l *уменьшение функционала* F *максимально. Среди всевозможных переносов такого рода.*

- *В результате разбиение* $\tilde{G}_{l+1}$ *отличается от* $\tilde{G}_l$ *только составом кластеров с номерами* μ и ν . Процесс
- завершается, когда никакой последующий перенос не уменьшает функционал или достигнуто указанное пользователем максимальное число итераций

Методы кластерного анализа

Пусть в результате применения разнообразных методов кластеризации получено множество различных решений для одних и тех же данных. При отсутствии внешнего критерия, выбор одного решения из данного множества кластеризаций может быть не ясен. Поэтому представляет интерес применение методов обработки полученных множеств кластеризаций с целью построения коллективных решений, более предпочтительных и обоснованных, чем полученные отдельными алгоритмами кластеризации.

Решение задачи кластерного анализа коллективами алгоритмов

Кластеризацию выборки $\tilde{S}_{ini} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, включающую кластеры $G_1, \ldots, G_k$ можно описать с помощью информационной матрицы $\| \alpha_{ji} \|_{m \times k}$, где $\alpha_{ji} = 1$, если $\mathbf{x}_j \in G_i$ и $\alpha_{ji} = 0$ в противном случае.

Наличие нескольких единиц в одной строке соответствует принадлежности объекта сразу нескольким кластерам. Нулевая строка означает отказ от кластеризации соответствующего объекта.

Решение задачи кластерного анализа коллективами алгоритмов

Определение 1. Информационные матрицы $I = \| \alpha_{ji} \|_{m \times k}$ и $I' = \| \alpha'_{ji} \|_{m \times k}$ называются эквивалентными, если они равны с точностью до перестановки столбцов.

Произвольная информационная матрица $I = \| \alpha_{ji} \|_{m \times k}$ определяет класс всех эквивалентных ей матриц $\tilde{K}(I)$.

Определение 2. Алгоритмом кластеризации A^c *называется* алгоритм, переводящий выборку $\tilde{S}_{ini}$ в класс эквивалентности $\tilde{K}(I)$ некотоой информационной матрицы I . .

Решение задачи кластерного анализа коллективами алгоритмов

Иными словами $A^c(\tilde{S}_{ini}) = \tilde{K}(\| \alpha_{ji} \|_{m \times k})$. Данное определение отражает факт свободы в обозначении полученных алгоритмом кластеров. Пусть существует r кластеризаций выборки $\tilde{S}_{ini}$ алгоритмами $A_1^c, \ldots, A_r^c$ на k кластеров. Задача построения оптимальной коллективной кластеризации состоит в вычислении по множеству из r *исходных* кластеризаций, задающих классы эквивалентности $\tilde{K}(\| \alpha_{ji}^1 \|_{m \times k}), \ldots, \tilde{K}(\| \alpha_{ji}^r \|_{m \times k})$ некоторого нового коллективного решения $\tilde{K}(\| \hat{\alpha}_{ji} \|_{m \times k})$, где $\hat{\alpha}_{ji} \in \{0,1\}$.

Решение задачи кластерного анализа коллективами алгоритмов

Оператор $\mathbf{B}(I_1,\ldots,I_r)=\|b_{ji}\|_{m\times k}$, $b_{ji}\in\{0,\ldots,r\}$, называется сумматором, если $b_{ji}=\sum_{t=1}^{r}\alpha_{ji}^{r}$

Матрицу, полученную в результате применения сумматора к некоторому набору информационных матриц, будем называть матрицей оценок. Оператор $\mathbf{C}$ называется решающим правилом, если $\mathbf{C}(\|b_{ji}\|_{m\times k})=\|\alpha_{ji}^{s}\|_{m\times k}$, где

$$\alpha_{ji}^{s}=\begin{cases}1 \text{ при } b_{ji}\geq b_{jt},\ t=1,\ldots,k\\ 0 \text{ в противном случае}\end{cases}\qquad j=1,\ldots,m$$

Решение задачи кластерного анализа коллективами алгоритмов

Определение 3. Комитетным синтезом информационной матрицы $\| \hat{\alpha}_{ji} \|_{m \times k}$ по множеству исходных кластеризаций, задаваемых набором инфориационными матриц $\tilde{I} = \{I_1, \ldots, I_r\}$, называется последовательное применение к $\tilde{I}$ сумматора $\mathbf{B}$ и решающего правила $\mathbf{C}$.

Для оценивания коллективного решения *вводится понятие* контрастных матриц оценок, соответствующих случаям, когда все исходные решения задач классификации оказались одинаковыми

Решение задачи кластерного анализа коллективами алгоритмов

Очевидно, что для произвольной контрастной матрицы $\| b_{ji}^c \|_{m\times k}$ $b_{ji}^c \in \{0, r\}$. Пусть $\tilde{B}^c = \{\| b_{ji}^c \|_{m\times k}\}$ - множество всевозможных контрастных матриц . Качеством произвольной матрицы $B = \| b_{ji} \|_{m\times k}$, вычисленной сумматором, определяется как минимальное расстояние до матриц из множества $\tilde{B}^c$:

$$\Phi(B) = \min_{B^c \in \tilde{B}^c} \sum_{j=1}^{m} \sum_{i=1}^{k} | b_{ji}^c - b_{ji} | \quad .$$

Решение задачи кластерного анализа коллективами алгоритмов

Набор информационных матриц $\{I'_1,\ldots,I'_r\}$ назовём эквивалентным $\{I_1,\ldots,I_r\}$, если $I'_1 \in \tilde{K}(I_1),\ldots,I'_r \in \tilde{K}(I_r)$

Задача оптимального коллективного синтеза сводится поиску эквивалентного $\{I_1,\ldots,I_r\}$ набора $\{I_1^m,\ldots,I_r^m\}$, для которого в результате применения сумматора получается матрица B_m с минимальным значением Φ среди всевозможных матриц, вычисляемых сумматором по наборам, эквивалентным $\{I_1,\ldots,I_r\}$.

Визуализация многомерных данных

При решении задач распознавания, классификации и анализа данных важное значение имеет наличие средств визуализации многомерных данных, позволяющих наглядно получать представление о конфигурации классов, кластеров и расположении отдельных объектов.

Пусть в n - мерном евклидовом пространстве задан набор из m элементов $\mathbf{x}_i \in \mathbf{R}^n, i = 1, \ldots, m$.

Визуализация многомерных данных

Требуется найти отображение этого набора точек на плоскость $\mathbf{R}^2$ *так,* чтобы метрические соотношения между образами точек на плоскости максимально соответствовали бы метрическим соотношениям между ними в исходном n -мерном признаковом пространстве: «близкие» («далекие») n - мерные точки, остались бы «близкими» («далекими») на плоскости.

Пусть $\mathbf{y}_i$ - отображение $\mathbf{x}_i$ на $\mathbf{R}^2$. δ_{ij} – расстояние между элементами $\mathbf{x}_i$ и $\mathbf{x}_j$ в $\mathbf{R}^n$, d_{ij} – расстояние между элементами $\mathbf{y}_i$ и $\mathbf{y}_j$ в $\mathbf{R}^2$

Визуализация многомерных данных

Ищется такое отображение, для которого сумма различий расстояний между точками будет минимальна

$$\mathrm{J}(\tilde{y}) = \sum_{j=1}^{m}\sum_{i=1}^{m}(d_{ij} - \delta_{ij})^2 \to \min,$$ где $$\tilde{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$$

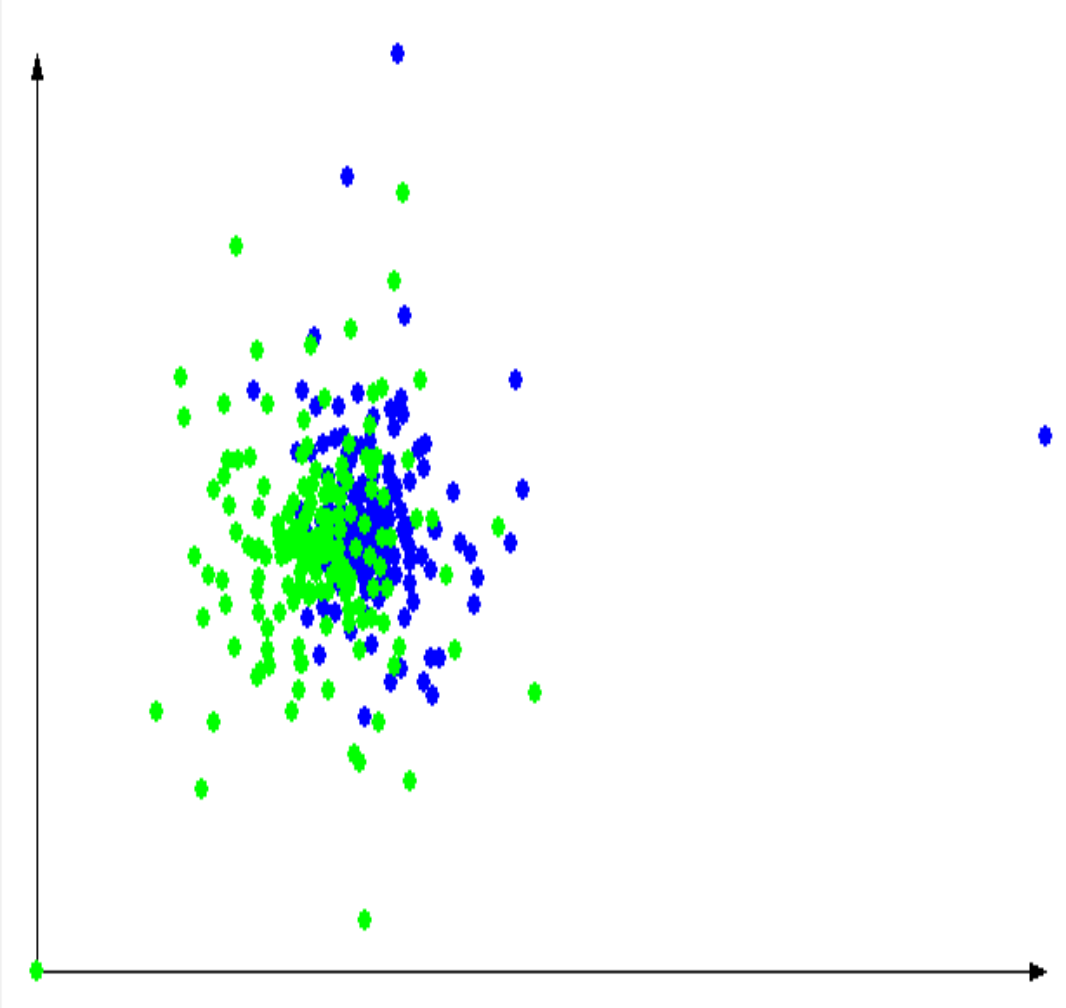
Минимизация функционала $\mathrm{J}(\tilde{y})$ проводится с помощью стандартной процедуры градиентного спуска

Визуализация многомерных данных

$\tilde{y}^{l+1} = \tilde{y}^{l} + \kappa * \mathbf{grad}[\mathrm{J}(\tilde{y}^{l})]$, где l - номер итерации, $\tilde{y}^{l}$ - конфигурация на плоскости, полученная на l -ой итерации, $\mathbf{grad}[\mathrm{J}(\tilde{y}^{l})]$ - градиент в “точке” $\tilde{y}^{l}$, κ - шаг градиентного спуска.

В качестве начальной конфигурации может использоваться проекция точек $\mathbf{x}_i \in \mathbf{R}^n, i = 1,\ldots,m$ на некоторую плоскость, соответствующую паре признаков.

Визуализация многомерных данных



Пример проекции на плоскость из пространства размерности 26, полученной описанным методом. Точкам зелёного и синего цвета соответствуют описания двух типов изображений.

Методы преобразования признакового пространства

Описанный метод многомерной визуализации фактически является методом нелинейного преобразования исходного признакового пространства. Вместе с тем существует эффективный метод линейной трансформации признакового пространства, позволяющий получить существенную информацию о существующих тенденциях. Данный метод называется Методом главных компонент (Principal component analysis) , а также преобразованием Карунена-Лоэв (Karhunen–Loeve transform)

Метод главных компонент

Метод главных компонент основан на переходе от исходного множества вообще говоря коррелированных переменных $X_1,\ldots,X_n$ к новому набору переменных $Y_1,\ldots,Y_n$ таких, что $\operatorname{cov}(Y_i,Y_j)=0 \text{ при } i\neq j,\ i,j=1,\ldots,n$

Переход к некоррелированным переменным может быть осуществлён с помощью линейного преобразования

,

$$Y_j=\sum_{j=1}^{n} w_{ij}X_i$$

Метод главных компонент

задаваемого матрицей вещественных коэффициентов

$\mathbf{W} = \| w_{ij} \|_{n \times n}$. Из симметричности ковариационной матрицы

$\| \mathrm{cov}(X_i, X_j) \|_{n \times n}$ следует, что

строки матрицы $\mathbf{W}$ на самом деле могут являются собственными векторами $\| \mathrm{cov}(X_i, X_j) \|_{n \times n}$.

При этом строки матрицы $\mathbf{W}$, соответствующие различным собственным значениям, всегда ортогональны друг другу, то есть $\sum_{j=1}^{n} w_{ij} w_{ji'} = 0 \; при \; i' \neq i, \; i, i' = 1, \ldots, n$

Метод главных компонент

Нетрудно показать также, что собственное значения матрицы $\| \operatorname{cov}(X_i, X_j) \|_{n \times n}$, соответствующее строке $(w_{1j}, \ldots, w_{nj})$ является дисперсией новой переменной $Y_j = \sum_{j=1}^{n} w_{ij} X_i$

Полученные в результате преобразования $\mathbf{W}$ переменные называют главными компонентами. Главные компоненты ранжируются в зависимости от величин соответствующих собственных значений.

Метод главных компонент

Переменная, соответствующая максимальному собственному значению Λ_1 и задаваемая соответствующим собственному вектором $\mathbf{w}^1$ называется первой главной компонентой. Она обладает максимальной дисперсией, равной Λ_1 . Следует отметить, что для гиперплоскости P_1, задаваемой направлением $\mathbf{w}^1$ и проходящей через геометрический центр $\tilde{S}$ достигает своего минимума сумма квадратов расстояний точек $\tilde{S}$ в пространстве исходных переменных от гиперплоскости P_1 .

Метод главных компонент

Вычтем из векторов - описаний объектов $\tilde{S}$ соответствующие значения первой компоненты. Для второй по величине собственного значения (дисперсии) Λ_2 главной компоненте, достигает минимума квадрат расстояния от изменённых векторов $\tilde{S}$ до гиперплоскости, задаваемой соответствующим Λ_2 собственным вектором $\mathbf{w}^2$ и проходящей через геометрический центр $\tilde{S}$.

Аналогичные свойства справедливы также и до последующих главных компонент.

Метод главных компонент

Интересно, что сумма квадратов отклонений для некоторой гиперплоскости пропорциональна величине соответствующего данной гиперплоскости собственного значения.

Метод главных компонент имеет широкий спектр применений, включая формирование нового признакового пространства, снижение размерности, визуализацию, анализ тенденций существующих в данных и др.