КЛАССИФИКАЦИИ ВУЗОВ ЧИТИНСКОЙ ОБЛАСТИ И АБАО НА ОСНОВЕ КЛАСТЕРНОГО АНАЛИЗА

Сайфутдинова А.С.

БГУЭП

Чита, Россия

Рассмотрим следующую задачу. Требуется оценить уровень образования в различных Вузах. Так как уровень образования – это понятие достаточно абстрактное, то получить его точную количественную характеристику практически невозможно. Однако его можно косвенно оценить по ряду экономических показателей, характеризующих стоимость обучения, квалификацию преподавательского персонала, материально-техническую базу, имидж Вуза и т.д. Очень часто далее строят некоторый интегральный показатель, объединяющий в себе все частные показатели, и на его основе ранжируют объекты (в нашем случае Вузы) по уровню образования. Однако такой подход имеет два основных недостатка:

1. Возможность компенсации низких значений одних показателей высокими значениями других. К примеру, если интегральный показатель равен простой сумме показателей, то Вуз, у которого стоимость обучения оценивается на 5, а качество подготовки на 3 будет эквивалентен Вузу, у которого стоимость обучения оценивается на 3, а качество обучения на 5. Это очевидно является абсурдным.
2. Возможность наличия сильной корреляционной зависимости между показателями, что искажает получаемые результаты.

Таким образом, прямое измерение уровня образования с помощью интегрального показателя представляется нецелесообразным. Альтернативу этому способу составляет кластерный анализ, являющийся одним из способов многомерной классификации, который не измеряет уровень образования, но позволяет сформировать группы относительно однородных Вузов, которые экспертным путем можно будет в дальнейшем охарактеризовать как группы ВУзов соответственно с очень высоким, высоким, средним, низким и очень низким уровнем образования.

Итак, имеется совокупность n объектов, каждый из которых характеризуется по k замеренным на нем признакам. Требуется разбить эту совокупность на однородные в некотором смысле группы (классы). При этом практически отсутствует априорная информация о характере распределения измерений внутри классов.

Полученные в результате разбиения группы обычно называются кластерами, а также таксонами или образами. Методы нахождения кластеров называются кластерным анализом.

В задачах кластерного анализа обычной формой представления исходных данных служит прямоугольная таблица:

$$X = \begin{pmatrix} x_{11} & x_{12} & \mathbf{K} & x_{1k} \\ x_{21} & x_{22} & \mathbf{K} & x_{2k} \\ \mathbf{K} & \mathbf{K} & \mathbf{K} & \mathbf{K} \\ x_{n1} & x_{n2} & \mathbf{K} & x_{nk} \end{pmatrix}$$

где x_{ij} - результат измерения j-го признака на i-ом объекте.

Наиболее трудным и наименее формализованным в задаче классификации является определение понятия однородности объектов. В общем случае понятие однородности объектов задается введением правила вычислений расстояния $r(x_i, x_j)$ между любой парой исследуемых объектов. Близкие с точки зрения этой метрики объекты считаются однородными, принадлежащими одному классу. При этом необходимо сопоставлять полученные расстояния с некоторым пороговым значением, определяемым в каждом конкретном случае по-своему.

Рассмотрим наиболее часто используемые расстояния в задачах кластерного анализа.

- ***Обычное евклидово расстояние***

$$r(x_i, x_j) = \sqrt{\sum_{p=1}^{k} (x_{ip} - x_{jp})^2}$$

где x_{ip}, x_{jp} - величина p-ой компоненты у i-го (j-го) объекта ($p = 1,...,k; i, j = 1,...,n$).

Естественно с геометрической точки зрения и содержательной интерпретации евклидово расстояние может оказаться бессмысленным, если его признаки имеют разные единицы измерения. Для приведения признаков к одинаковым единицам прибегают к нормировке каждого признака путем деления центрированной величины на среднее квадратическое отклонение и переходят от матрицы X к нормированной матрице с элементами

$$x_{ij}^{H} = \frac{x_{ip} - \bar{x}_p}{s_p},$$

где x_{ip} - значение p-го признака у i-го объекта; $\bar{x}_p$ - среднее арифметическое значение p-го признака;

$$s_p = \sqrt{\frac{1}{n}\sum_i \left(x_{ip} - \bar{x}_p\right)^2}$$

- среднее квадратическое отклонение p-го признака.

- «Взвешенное» евклидово расстояние

$$r\left(x_i, x_j\right) = \sqrt{\sum_{p=1}^{k} w_p \left(x_{ip} - x_{jp}\right)^2}$$

применяется в случаях, когда каждой компоненте x_p удается приписать некоторый «вес» w_p, пропорциональный степени важности признака в задаче классификации. Обычно принимают $0 \le w_p \le 1$, где p=1,...,k.

- *Хеммингово расстояние*

$$r\left(x_i, x_j\right) = \sum_{p=1}^{k} \left|x_{ip} - x_{jp}\right|$$

используется как мера различия объектов, задаваемых дихотомическими признаками, т.е. признаками, значения которых равны или 0, или 1. Хеммингово расстояние равно числу несовпадений значений соответствующих признаков в рассматриваемых объектах.

По мере того, как объекты объединяются в классы возникает необходимость измерения расстояния между этими классами. Наиболее употребительными расстояниями между классами объектов или кластерами являются:

1. расстояние, измеряемое по принципу «ближайшего соседа», т.е. расстояние между двумя ближайшими точками кластеров

$$r\left(S_l, S_m\right) = \min r\left(x_i, x_m\right)$$

2. расстояние, измеряемое по принципу «дальнего соседа», т.е. расстояние между двумя самыми дальними точками кластеров

$$r\left(S_l, S_m\right) = \max r\left(x_i, x_m\right)$$

3. расстояние, измеряемое по «центрам тяжести» групп

$$r\left(S_l, S_m\right) = r\left(\bar{x}_l, \bar{x}_m\right)$$

4. расстояние, измеряемое по принципу «средней связи» (Это расстояние определяется как среднее арифметическое всех попарных расстояний между представителями рассматриваемых групп)

$$r\left(S_l, S_m\right) = \frac{1}{n_l n_m} \sum_{x_i \in S_l} \sum_{x_j \in S_m} r\left(x_i, x_j\right)$$

Так как существует большое количество различных способов разбиения на классы заданной совокупности элементов, то представляет интерес задача сравнительного анализа качества этих способов разбиения. С этой целью вводится понятие функционала качества разбиения Q(S), определенного на множестве всех возможных разбиений.

Существуют следующие виды функционала качества:

1. сумма внутриклассовых дисперсий

$$Q_1(S) = \sum_{m=1}^{p} \sum_{x_i \in S_m} r^2(x_i, \bar{x}_m)$$

2. сумма попарных внутриклассовых расстояний между элементами

$$Q_2(S) = \sum_{m=1}^{p} \sum_{x_i, x_j \in S_m} r^2(x_i, x_j)$$

Иерархические кластер-процедуры

Иерархические (деревообразные) процедуры являются наиболее распространенными алгоритмами кластерного анализа. Они бывают двух типов: агломеративные и дивизимные.

Принцип работы иерархических агломеративных процедур состоит в последовательном объединении групп элементов сначала самых близких, а затем все более отдаленных друг от друга.

Принцип работы иерархических дивизимных процедур состоит в последовательном разделении групп элементов сначала самых далеких, а затем все более близких друг к другу.

Результаты

Поскольку кластерный анализ позволяет находить расстояние между объектами по любому количеству показателей, то целесообразной будет организация выбора их состава.

Расстояние между кластерами предлагается находить тремя способами. Это метод ближнего соседа, метод дальнего соседа и метод среднего значения.

В результате мы получим таблицу, в которой для каждого Вуза по указанным факторам будут даны оценки 6 экспертов, усредненные нами по весовым коэффициентам.

В результате получается ряд таблиц, представляющих собой все более укрупненное объединение Вузов в кластеры. Последней мы получаем таблицу 2 на 2, в которой все объекты разбиты на 2 кластера. В зависимости от цели исследования выбирается то или иное количество кластеров. Соответственно получается несколько групп однотипных Вузов. Для каждой группы необходимо разработать соответствующие рекомендации, отвечающие цели исследования.

Для сравнения с результатами, полученными при построении матрицы Мак-Кинси, рассмотрим в качестве факторов интегрированные показатели привлекательности и конкурентоспособности.

Кластерный анализ позволяет нам разделить объекты на требуемое количество групп вне зависимости от количества имеющихся у нас показателей, легко исключить ненужные показатели или связанные друг с другом, но для интерпретации результата, характеристики каждой группы необходимо применение каких-то других методов, чаще всего экспертных оценок. Именно поэтому матричный метод, в частности, построение матрицы Мак-Кинси, столь удобны, они позволяют не только разбить на группы (на 9 групп), но и получить наглядную характеристику объектов, попавших в ту или иную группу.