Концептуальные основы формирования кластеров на примере рыбоводческих организаций в Удмуртской Республике

© 2012 О.В. Кузнецова
Ижевская государственная сельскохозяйственная академия
E-mail: kuzov@list.ru

В статье рассмотрены методы и алгоритмы кластеризации объектов с разнородными характеристиками. На их основе выявлены и обоснованы кластеры организаций рыбоводства в Удмуртской Республике.

Ключевые слова: кластер, методы кластеризации, рыбоводство.

Как известно, статистическая обработка информации предполагает создание однородной информационной базы. Но нередко бывает так, что требуется проанализировать множество объектов, характеризующихся разнородными показателями, например, 28 рыбоводческих предприятий на территории Удмуртской Республики, которые имеют разные результаты хозяйственной деятельности, организационно-правовые формы, функционируют в разных природно-климатических условиях и условиях инфраструктурной обеспеченности.

Для классификации упомянутых рыбоводческих предприятий мы применили один из методов многомерного анализа неоднородных статистических совокупностей - кластерный анализ. Метод основан на том, что результаты отдельных наблюдений представляются в виде точек некоторого многомерного геометрического пространства и затем объединяются в группы как “сгустки” этих точек[1]. В нашем случае точками являются рыбоводческие хозяйства (28 точек), а их координатами - значения показателей. Мы проанализировали предприятия по девяти признакам, т.е. каждая точка имеет 9 координат.

Целью такого анализа выступает группировка исходных многомерных данных так, чтобы элементы внутри этих групп (кластеров) были максимально близки (похожи), а элементы из разных групп - максимально далеки друг от друга (не похожи).

Итак, мы проанализировали 28 предприятий Удмуртской Республики, занимающихся разведением товарной рыбы (карп - в основном, осетр, форель), по показателям, представленным в табл. 1. Ранее нами был разработан алгоритм вычисления интегрального показателя инфраструктурной обеспеченности рыбоводческого хозяйства, по которому каждому из 28 предприятий было приведено в соответствие некоторое значение этого показателя[2].

Существует множество алгоритмов кластерного анализа, отличающихся не только формулами, применяемыми при вычислениях, но и концепциями. Подробнее с каждым из них можно ознакомиться в[3].

Опишем основные этапы анализа.

1) Нормализация значений показателя. Процесс нормализации (нормировки) превращает показатели, изначально выражающиеся в разных единицах (руб., т, шт., % и т.д.) в безразмерные величины. В итоге получаем так называемую новую условную единицу измерения, допуска-

Таблица 1. **Показатели рыбоводческих предприятий Удмуртской Республики**

№ п/п	Показатели	Предприятие № 1	Предприятие № 2	…	Предприятие № 28
1	Интегральный показатель инфраструктуры, баллов	x_{11}	x_{12}	…	x_{128}
2	Выход продукции (товарная рыба), т	x_{21}	x_{22}	…	x_{228}
3	Выход продукции (рыбопосадочный материал), тыс. шт.	x_{31}	x_{32}	…	x_{328}
4	Себестоимость единицы (ц) продукции, руб.-коп.	x_{41}	x_{42}	…	x_{428}
5	Объем реализованной рыбы, т	x_{51}	x_{52}	…	x_{528}
6	Рыбопродуктивность, ц/га	x_{61}	x_{62}	…	x_{628}
7	Площадь водного зеркала, га	x_{71}	x_{72}	…	x_{728}
8	Средняя заработная плата работников, руб.	x_{81}	x_{82}	…	x_{828}
9	Коэффициент специализации	x_{91}	x_{92}	…	x_{928}

Примечание. x_{ij} – значение i-го показателя для j-го хозяйства.

ющую формальное сопоставление объектов. Нормализацию данных обычно проводят по одной из следующих формул[4]:

$$z = \frac{x - \bar{x}}{\sigma}; \quad (1)$$

$$z = \frac{x}{\bar{x}}; \quad (2)$$

$$z = \frac{x}{x'}; \quad (3)$$

$$z = \frac{x}{x_{\max}}; \quad (4)$$

$$z = \frac{x - \bar{x}}{x_{\max} - x_{\min}}, \quad (5)$$

где $\bar{x}$ - среднее значение признака;

σ - среднее квадратическое отклонение признака;

$x2$ - некоторое эталонное (нормативное) значение признака;

$x_{\min}$ и $x_{\max}$ - соответственно, минимальное и максимальное значения признака.

Существуют и другие способы нормировки данных, но они обычно являются вариациями приведенных выше. Нами был применен первый способ - наиболее популярный из способов, его еще называют выравнивающей дисперсией всех признаков.

Итак, вычисляем среднее значение $\bar{x}_i (i = 1...9)$ и среднее квадратическое отклонение σ_i по каждому признаку, затем по формуле (1) определяем нормализованные значения данных (поскольку процедура кластеризации достаточно трудоемка, все расчеты производим при помощи компьютера в программе Excel). В итоге получаем таблицу чисел из 9 строк (количество признаков) и 28 столбцов (количество предприятий) (табл. 2).

Таблица 2. **Нормализованные значения исходных данных**

№	1	2	…	28
1	z_{11}	z_{12}	…	z_{128}
2	z_{21}	z_{22}	…	z_{228}
3	z_{31}	z_{32}	…	z_{328}
4	z_{41}	z_{42}	…	z_{428}
5	z_{51}	z_{52}	…	z_{528}
6	z_{61}	z_{62}	…	z_{628}
7	z_{71}	z_{72}	…	z_{728}
8	z_{81}	z_{82}	…	z_{828}
9	z_{91}	z_{92}	…	z_{928}

2) При образовании кластеров неизбежно встает вопрос об измерении расстояния между точками (объектами). Измерить расстояние между объектами (определить степень близости) также можно различными способами: вычислить линейное расстояние, Евклидово расстояние, обобщенное степенное расстояние Минковского и т.д. С подробным описанием каждого способа можно ознакомиться в[5]. Наиболее популярной метрикой в кластерном анализе является Евклидово расстояние:

$$\rho(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2},$$

где $(x_1, x_2, ..., x_n)$ и $(y_1, y_2, ..., y_n)$ - координаты точек x и y в n-мерном пространстве.

Итак, вычисляем расстояния между каждой парой точек (предприятий). В результате получаем 378 чисел - количество комбинаций (сочетаний) двух элементов из 28 элементов:

$$C_{28}^2 = \frac{28!}{2!26!} = \frac{26!27 \cdot 28}{2 \cdot 26!} = 27 \cdot 14 = 378.$$

Полученные числа удобно записать также в виде таблицы расстояний (табл. 3).

Таблица 3. **Расстояния между предприятиями**

№	1	2	3	…	28
1	0	ρ_{12}	ρ_{13}	…	ρ_{128}
2		0	ρ_{23}	…	ρ_{228}
3			0	…	ρ_{328}
…				…	…
27				0	ρ_{2728}
28					0

Примечание:

$\rho_{ij} = \sqrt{(z_{1i} - z_{1j})^2 + z_{2i} - z_{2j})^2 + ... + ... + z_{9i} - z_{9j})^2}$ - расстояние между i-й и j-й точками.

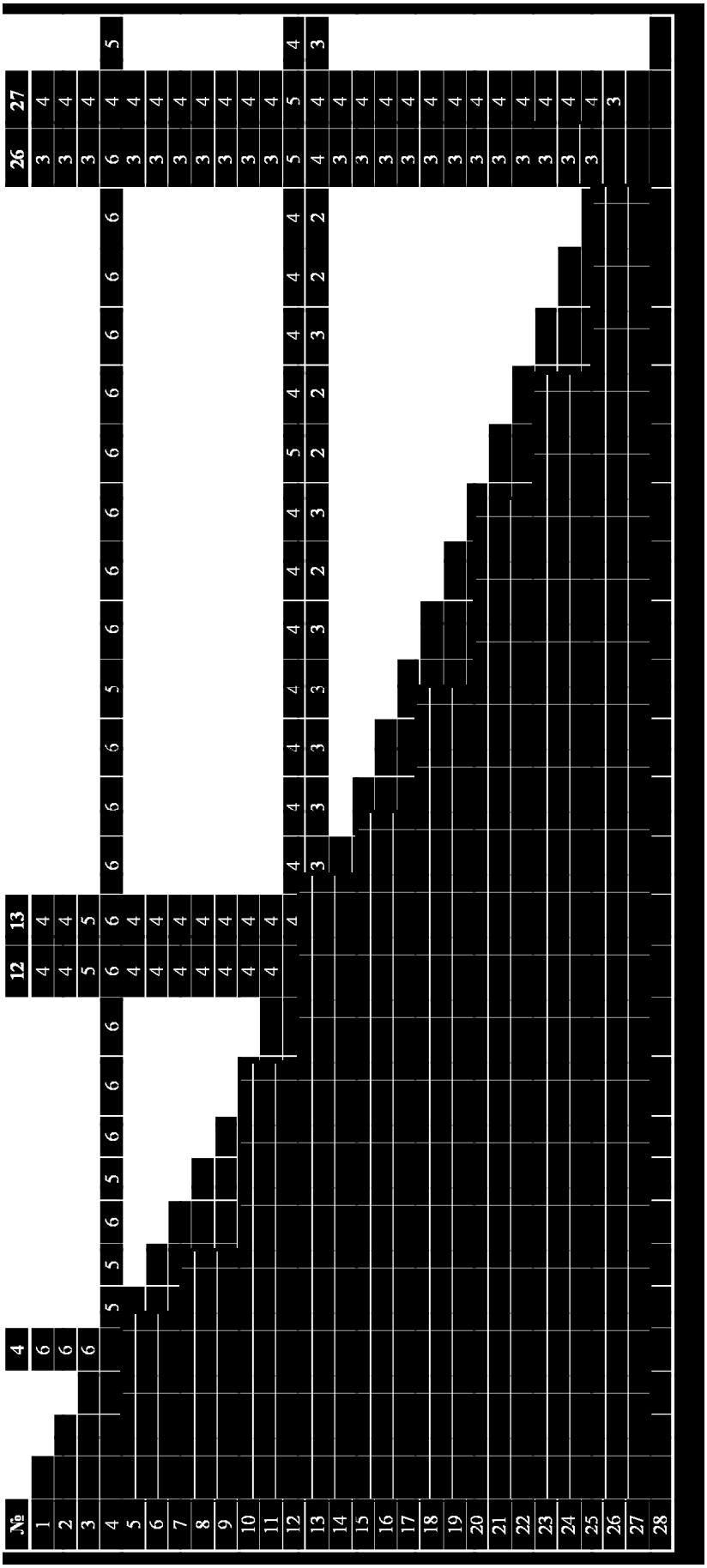
3) Следующий этап - образование кластеров. В литературе рассматривается около семи десятков алгоритмов кластеризации (иерархические алгоритмы, алгоритмы упорядочения (диагонализации) матрицы расстояний, алгоритмы эталонного типа, разрезания графа, прочие и комбинированные), не лишенных субъективности[6]. Как говорит И.Д. Мендель, “методы визуализации... внутренне парадоксальны - они используют точные алгоритмы... лишь для того, чтобы впоследствии человек принял на их основе весьма приближенное, естественное в его понимании решение. Однако такая парадоксальность лежит в природе вещей и не тормозит познание, а способствует его успехам”[7].

Для образования кластеров мы применили метод группировки:

а) определили минимальное и максимальное из полученных расстояний: $\rho_{\min}$ и $\rho_{\max}$;

б) отрезок изменения расстояний поделили на части (частичные интервалы) одинаковой длины.

Таблица 4. Расстояния между предприятиями (по принадлежности интервалам)



$h = \frac{\rho_{max} - \rho_{min}}{k}$ - длина частичных интервалов.

Количество частичных интервалов определили по формуле: k=1+3,322lgn. В нашем случае n=28, тогда $k \approx 6$;

в) затем каждому значению ρ_{ij} присвоили номер интервала (от 1 до 6) в зависимости от того, какому интервалу оно принадлежит.

В результате таблица расстояний приобрела следующий вид (табл. 4). Большей цифре в ячейках таблицы соответствует большее расстояние между предприятиями.

Теперь, глядя на полученные цифры, объединяем элементы в группы (кластеры). Поскольку элементы 1, 2, 3, 5-11, 14-25 и 28 находятся друг от друга на наименьших расстояниях, объединяем их в кластер 1. Сюда вошло большинство крестьянско-фермерских хозяйств и индивидуальных предпринимателей, а также средние предприятия, в которых производство товарной рыбы сочетается с другими видами деятельности. Предприятие 4 находится на большом расстоянии от всех остальных предприятий. Следовательно, ни с одним из них образовать кластер оно не может. Таким образом, получаем кластер 2, состоящий из единственного предприятия 4 СГУП "Рыбхоз "Пихтовка"". Элемент 12 (ООО "Каракулинский рыбхоз "Прикамье"") также отдален от всех других, относим его к отдельному кластеру 3. Элемент 13 (ООО "Русь") к кластеру 1 отнести не можем, поскольку, хотя он находится и близко к некоторым точкам этого кластера, но там присутствуют в числе прочих и далеко расположенные от него элементы, за счет чего внутрикластерное расстояние может существенно увеличиться. Следовательно, выделяем его в отдельный кластер - 4. Рассуждая аналогично, образуем единичные кластеры 5 (из элемента 26 - ООО "Рыбоводный модуль") и 6 (из элемента 27 - ООО "Аквафонд"). Следует заметить, что элементы 26 и 27 можно было бы объединить, но дальнейшая проверка показала, что главное условие кластеризации в этом случае не выполняется.

4) Далее проводим процедуру проверки правильности образования кластеров. Целью проведенного анализа было образование наиболее удаленных друг от друга групп, состоящих из наиболее схожих элементов. Значит, расстояния между точками каждого кластера должны быть меньше расстояний от этого кластера до всех остальных.

Следующим шагом было определение внутрикластерных и межкластерных расстояний. Из расстояний между предприятиями (табл. 5), входящими в определенный кластер, выбираем наибольшее; если кластер единичный, то это расстояние равно нулю.

Таблица 5. **Внутрикластерные и межкластерные расстояния**

№ кластера	2	3	4	5	6
1	6,315	6,096	5,819	5,921	5,674
2		4,461	4,495	4,499	3,637
3			3,654	4,213	3,737
4				2,249	3,362
5					2,471

Расстояние между кластерами вычисляем аналогично расстоянию между элементами (Евклидово расстояние). Точками в этом случае являются кластеры, а их координаты равны средним значениям соответствующих координат элементов, принадлежащих кластеру. Результаты этапа изложены в табл. 5.

Максимальное внутреннее расстояние в кластере 1 - 3,521, что существенно меньше, чем расстояния между кластером 1 и всеми остальными. Оставшиеся кластеры единичные, следовательно, для них основной критерий кластеризации также выполняется.

Практическое применение проведенной кластеризации рыбоводческих предприятий состоит в том, что к разным кластерам необходимо применять разные методы управления и оказывать разные формы государственной поддержки (субсидии и т.д.).

---

[1] *Мандель И.Д.* Кластерный анализ. М., 1988. С. 4.

[2] *Алексеева Н.А., Кузнецова О.В.* Оценка инфраструктуры развития рыбного хозяйства в Удмуртской Республике // Перспективы науки. 2011. □ 9 (24).

[3] *Мандель И.Д.* Указ. соч. С. 36.

[4] Там же. С. 27.

[5] Там же. С. 31.

[6] См.: Там же. С. 53; *Степанов Р.Г.* Технология Data Mining: Интеллектуальный анализ данных / Каф. экон. кибернетики КГУ: сайт. URL: http://m8.ksu.ru/EOS/dm.pdf.

[7] *Мандель И.Д.* Указ. соч. С. 126.

Поступила в редакцию 05.12.2011 г.