34. Функционалы качества разбиения на классы в кластерном анализе.

1.5. Критерии качества классификации

Наиболее распространенными функционалами качества являются:

1. **Сумма квадратов расстояний до центров классов.** При использовании этого критерия стремится получить такое разбиение совокупности объектов на k кластеров, при котором значение этого функционала было бы минимальным, т.е.:

$$F_1 = \sum_{l=1}^{k} \sum_{i \in S_l} d^2(X_i, \overline{X_l}) \to \min,$$

(15)

где l номер класса; X_l - центр l-го класса; X_i - вектор значений переменных для i-го объекта, входящего в кластер S_l; $d(X_i, X_l)$ - расстояние между i-м объектом и центром класса S_l

Примечание 1: это внутриклассовая вариация - мера того, насколько объекты внутри одного кластера близки к центру

Примечание 2: минимизируем, чтобы получить компактные кластеры

2. Сумма внутриклассовых расстояний между объектами.

Наилучшим следует считать такое разбиение, при котором достигается минимальное значение этого функционала, т.е. получены области.

$$F_2 = \sum_{l=1}^{k} \sum_{i,j \in S_l} d_{ij}^2 \to \min,$$

(16)

где d_{ij} - расстояние между i-м и j-м объектами, входящих в кластер S_l.

Примечание 1: Это плотность кластеров — мера того, насколько близки друг к другу объекты внутри одного кластера.

Примечание 2: чем меньше все попарные расстояния между объектами в каждом кластере тем плотнее кластер

Сравнение с F_1:

- F_1: "каждый объект близок к центру"
- F_2: "все объекты близки друг к другу" (более строгое условие)

3. Сумма внутриклассовых дисперсий.

Оптимальным следует считать такое разбиение, при котором достигается миниманое значение этого функционала, т.е. получены однородные кластеры:

$$F_3 = \sum_{l=1}^{k} \sum_{j=1}^{p} \sigma_{ij}^2 \to \min,$$

(17)

где σ_{ij}^2 - дисперсия j-й переменной в кластере S_l.

На принципе минимизации внутриклассовой дисперсии основаны алгоритмы метода k-средних и метода Уорда.

Примечание: Это однородность кластеров по каждому признаку. (Стремимся к тому, чтобы внутри кластера объекты были похожи по всем характеристикам)

4. Средние межклассовые расстояния.

Оптимальным следует считать такое разбиение, при котором достигается максимальное значение этого функционала, т.е. чем дальше кластеры друг от друга, тем лучше:

$$F_3 = \frac{\sum_{i \in S_q} d_{ij}}{\sum_{n \in N_q} nq} \to \max .$$

,
где S_q - множество объектов в кластере q,
N_q — количество объектов в кластере q,
n - выборка, q - номер кластера

(18)

Примечание 1: Это разделимость кластеров — мера того, насколько хорошо кластеры отделены друг от друга.

Примечание 2: Максимизируем среднее расстояние между объектами из разных кластеров