Кластерный анализ

ОСНОВНЫЕ ВОПРОСЫ

Кластерный анализ (англ. *Data clustering*) — задача разбиения заданной выборки *объектов* (ситуаций) на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Кластер — группа элементов, характеризуемых общим свойством.

Главная цель кластерного анализа — нахождение групп схожих объектов в выборке

Примеры применения кластерного анализа:

- археология,
- медицина,
- психология,
- химия,
- биология,
- информационная безопасность,
- филология,
- антропология,
- социология и другие области.

Задачи КА

- Разработка типологии или классификации.
- Исследование полезных концептуальных схем группирования объектов.
- Порождение гипотез на основе исследования данных.
- Проверка гипотез или исследования для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных

Этапы КА

- Отбор выборки для кластеризации .
- Определение множества переменных, по которым будут оцениваться объекты в выборке.
- Вычисление значений той или иной меры сходства между объектами.
- Применение метода кластерного анализа для создания групп сходных объектов.
- Проверка достоверности результатов кластерного решения

Требования к данным

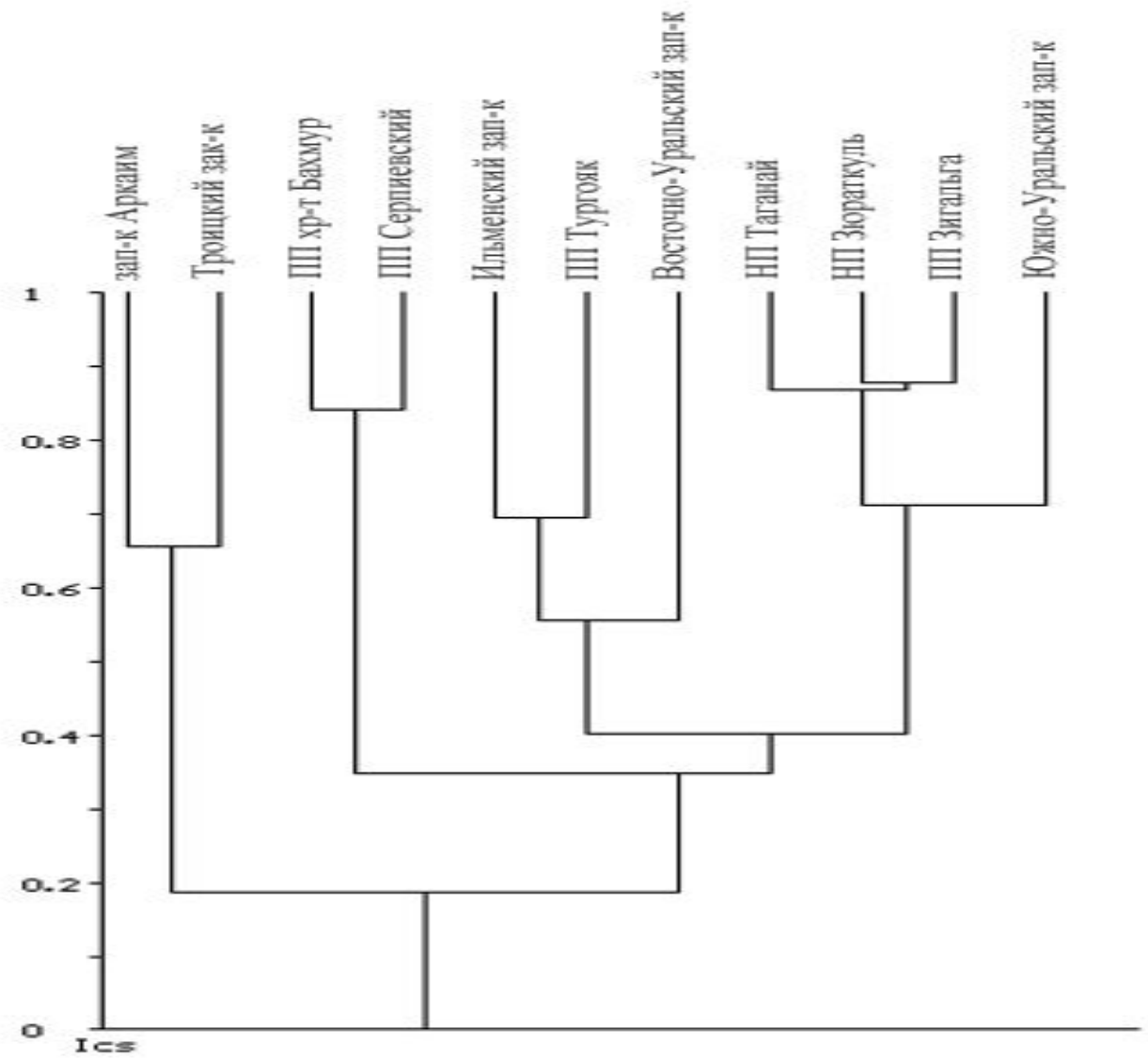
- Кластерный анализ предъявляет следующие *требования к данным*:
- показатели не должны коррелировать между собой
- показатели должны быть безразмерными
- распределение показателей должно быть близко к нормальному
- показатели должны отвечать требованию «устойчивости», под которой понимается отсутствие влияния на их значения случайных факторов
- выборка должна быть однородна, не содержать «выбросов»

Анализ и интерпретация результатов КА

При анализе результатов социологических исследований рекомендуется осуществлять анализ методом Уорда, при котором внутри кластеров оптимизируется минимальная дисперсия, в итоге создаются кластеры приблизительно равных размеров.

Метод Уорда наиболее удачен для анализа социологических данных. В качестве меры различия лучше квадратичное евклидово расстояние, которое способствует увеличению контрастности кластеров

ПРИМЕР ДЕНДРОГРАММЫ (СОСУЛЬЧАТОЙ ДИАГРАММЫ)- ОХРАНЯЕМЫЕ АРХЕОЛОГИЧЕСКИЕ ОБЪЕКТЫ ЧЕЛЯБИНСКОЙ ОБЛАСТИ



Методы кластеризации

- K-K-средних (K-средних (K-means)
- Иерархическая кластеризация Иерархическая кластеризация или таксономия
- Нейронная сеть Кохонена
- Алгоритмы семейства KRAB
- Статистические алгоритмы кластеризации
- Графовые алгоритмы кластеризации

Формальная постановка задачи кластеризации

Пусть X — множество объектов, Y — множество номеров (имён, меток) кластеров. Задана функция расстояния между объектами $\rho(x, x')$. Имеется конечная обучающая выборка объектов $X^m = \{x_1, \ldots, x_m\} \subset X$. Требуется разбить выборку на непересекающиеся подмножества, называемые *кластерами*, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера y_i.

Алгоритм кластеризации — это функция $a: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$. Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного *критерия качества* кластеризации .