Оценка обоснованности кластеризации для данных высокой размерности

Сивоголовко Елена Владимировна, гр. 544

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра системного программирования

Научный руководитель: к.ф.-м.н., профессор. Новиков Б.А.
Рецензент: ассистент кафедры информатики Васильева Н.С.

Санкт-Петербург
2008г.

Обоснованность кластеризации

Критерии обоснованности:

1. Компактность кластеров
2. Отделимость кластеров
3. Сосредоточеность элементов у центра кластера

Оценка обоснованности кластеризации для CBIR

- Применение кластеризации для поиска изображений по содержанию.
- Проблемы кластеризации изображений.
- Важность оценки обоснованности.

Метрики обоснованности

- Индекс Данна

$$D = \min_{i=1\ldots n_c} \left\{ \min_{j=i+1\ldots n_c} \left(\frac{d(c_i, c_j)}{\max_{k=1\ldots n_c} (diam(c_k))} \right) \right\} \quad (1)$$

- RMSSDT индекс

$$RMSSTD = \sqrt{\frac{\sum_{i=1}^{n_c} \sum_{j=1}^{d} \sum_{k=1}^{n_{ij}} (x_k - \bar{x}_j)^2}{\sum_{i=1}^{n_c} \sum_{j=1}^{d} (n_{ij} - 1)}} \quad (2)$$

Плотность кластерного ядра

- Ядро кластера

$$R_i = \bar{r}_i + \sqrt{2} * \sigma_i$$

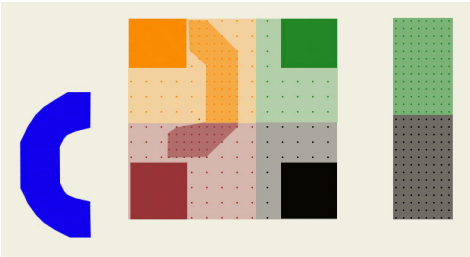
- Плотность ядра

$$\rho_i = \frac{R_i}{\overline{N_i}}$$

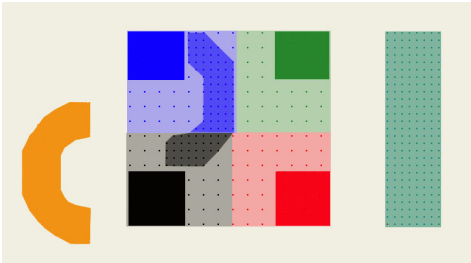
- Средняя плотность разбиения

$$\bar{\rho} = \frac{\rho_1 + \ldots + \rho_{n_c}}{n_c}$$

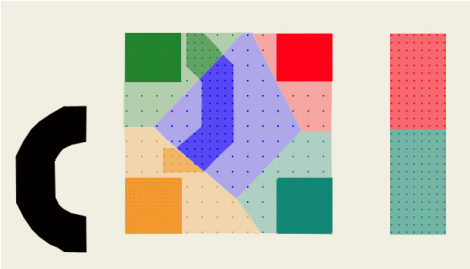
Синтетические 2D данные



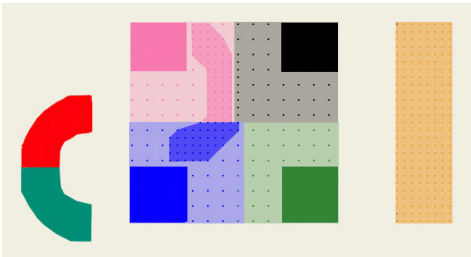
1. разбиение на пять кластеров



2. разбиение на шесть кластеров(верно)



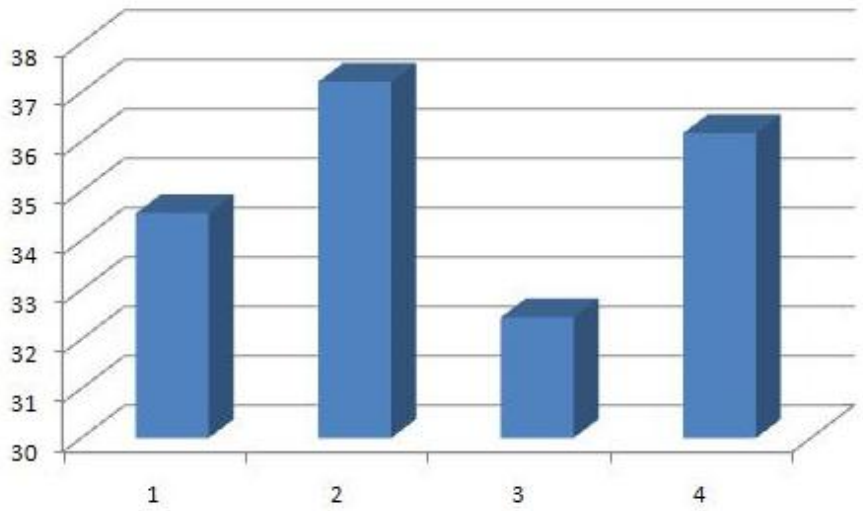
3.разбиение на шесть кластеров(неверно)



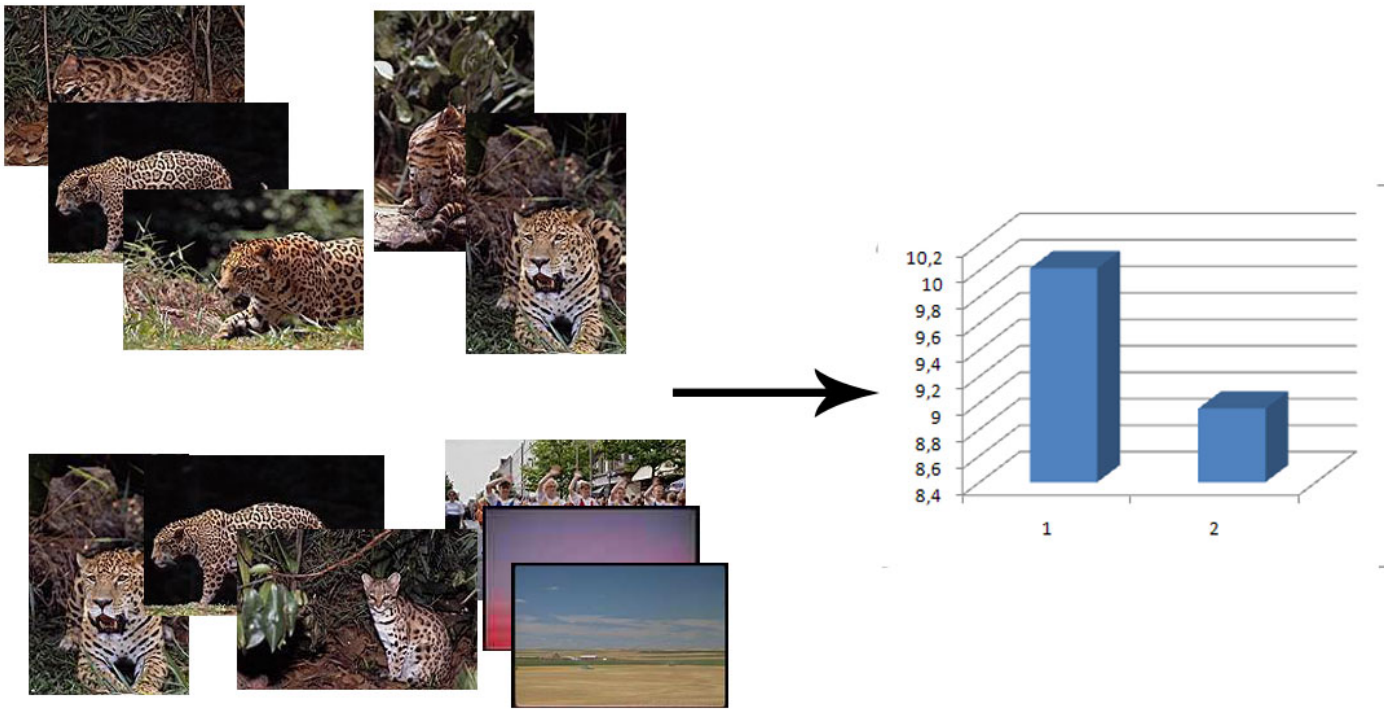
4. разбиение на семь кластеров

Показатели плотности для различных разбиений

средняя плотность ядра

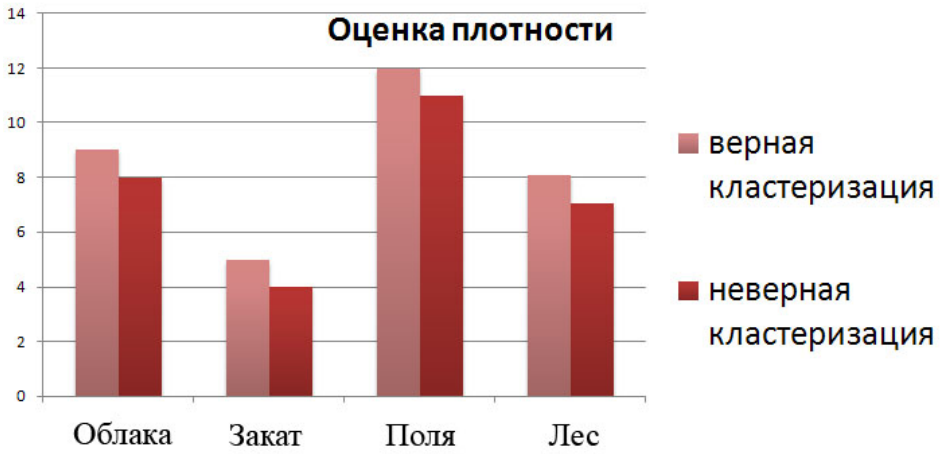


Межкластерное сравнение



База данных CorelSmall-100. Алгоритм кластеризации: сферический к-средних. Сравнение по семантическому классу "Кошки".

Межкластерное сравнение



Заключение

Введена новая метрика обоснованности кластеризации – плотность кластерного ядра, которая позволяет определить компактность кластеров и составить представление о их внутренней структуре.
Эффективность метрики подтверждается рядом экспериментов, как на синтетических, так и на реальных данных.

Дальнейшие направления исследований

Планируется обобщить метрику для следующих случаев:

- нечеткая кластеризация
- не гауссово распределение данных
- метрические пространства