Санкт-Петербургский Государственный Университет
Факультет Прикладной Математики – Процессов Управления

Исследование и применение кластеризации потоков данных

Дипломная робота студнта 64 группы:Ван Ячжо
Научный руководитель:
профессор, д.ф.-м.н. Зубов С.В.

Кластеризация

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

Цели кластеризации

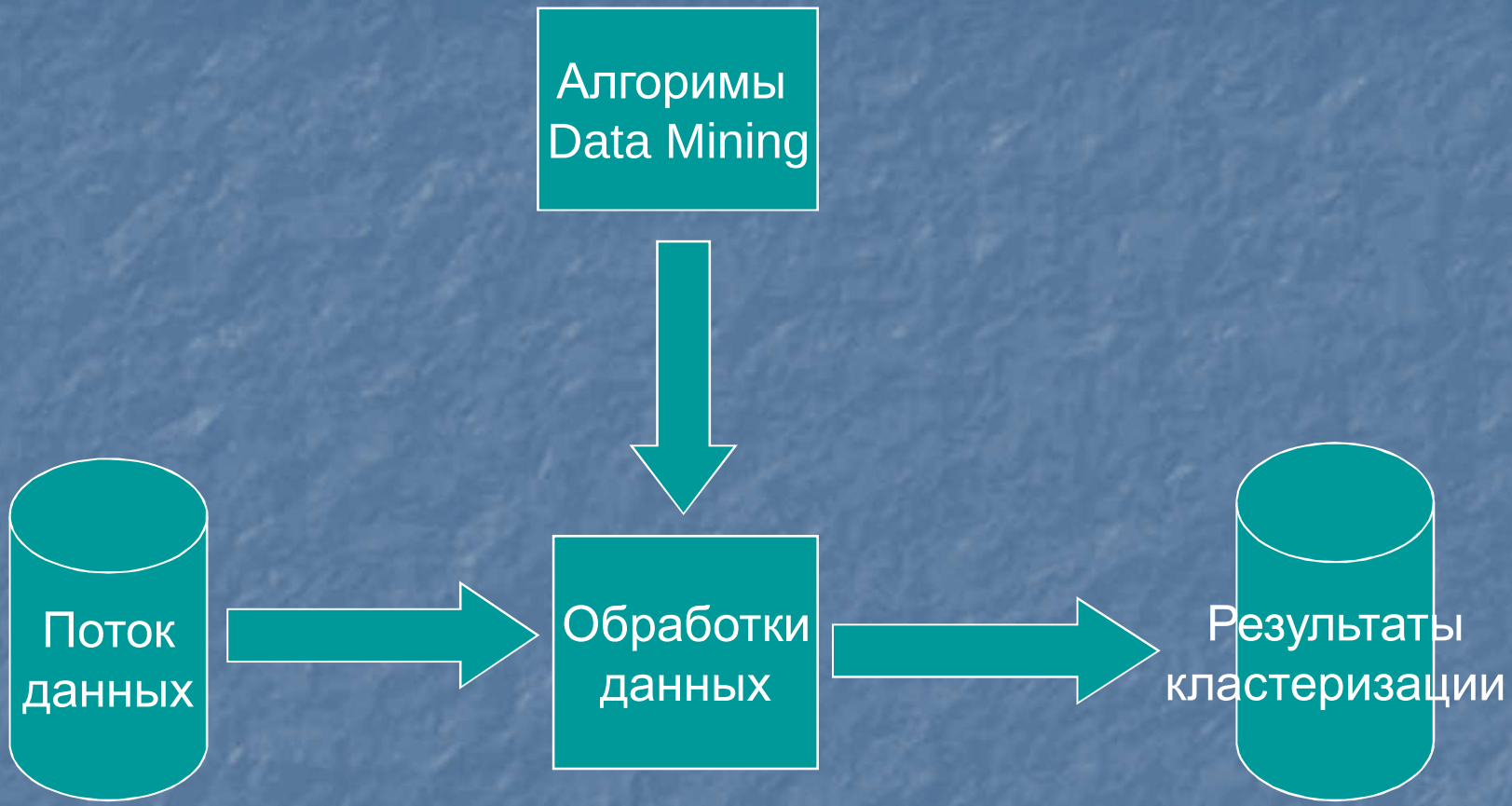
- Понимание данных путём выявления кластерной структуры.Разбиение выброки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятия решений
- Сжатие даных .Если исходная выброка избыточно большая , то можно сократить её , оставив по одному наиболее типичному представителю от каждого кластера.
- Обнаржение новизны . Выделяются нетипичные объекты , которые не удаётся присоединить ни к одному из кластеров

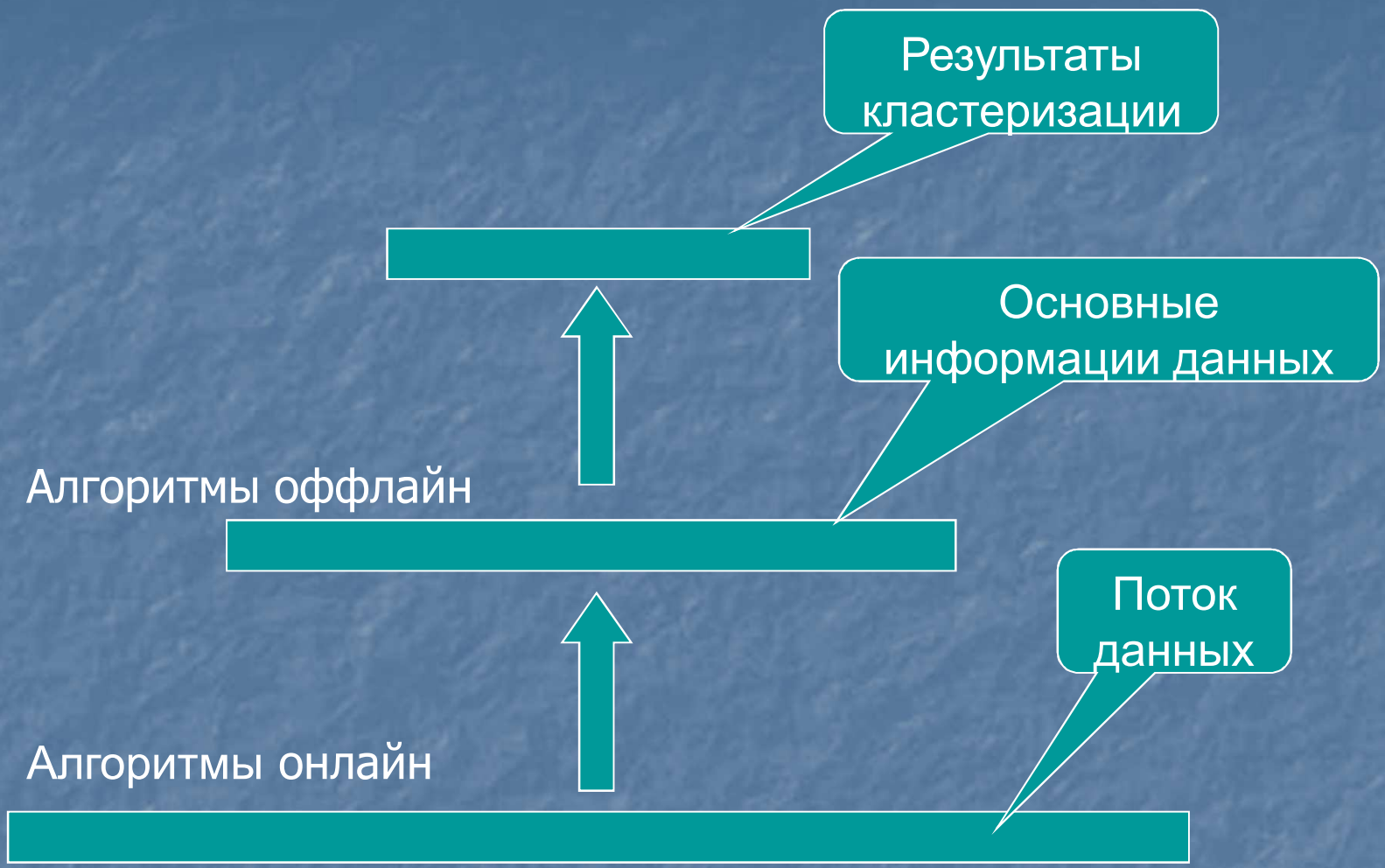
Методы кластеризации

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

Кластерный анализ потока данных на двух слое

- Алгоритмы онлайн обработает на данные грубый но быстрый , и сохраняет средние результаты;
- Алгоритмы оффлайн точно и сложно анализирует средние результаы , и полуичть конечные результаты.


Алгоримы
Data Mining
Поток
данных
Обработки
данных
Результаты
кластеризации


Результаты кластеризации
Основные информации данных
Алгоритмы оффлайн
Поток данных
Алгоритмы онлайн

Алгоритм онлайн

Sdmicro-cluster(W,D,Datastream)

W размер окна данных , D значение порога плотности в кластере

{Кластеризация потоки данных с помощью методов плотности.

While(поток данных не законится)

{for(i=0; i<W; i++)

{читать одну точку доступу в поток данных;

Опредерить эту точку принадлежает ли в сущесвующей кластерах.

if(существует такой кластер)

положить эту точку в кластер и поменять её собственный значение;

else

Создать новый кластер для той точки , и добавлять её в таблице hash;

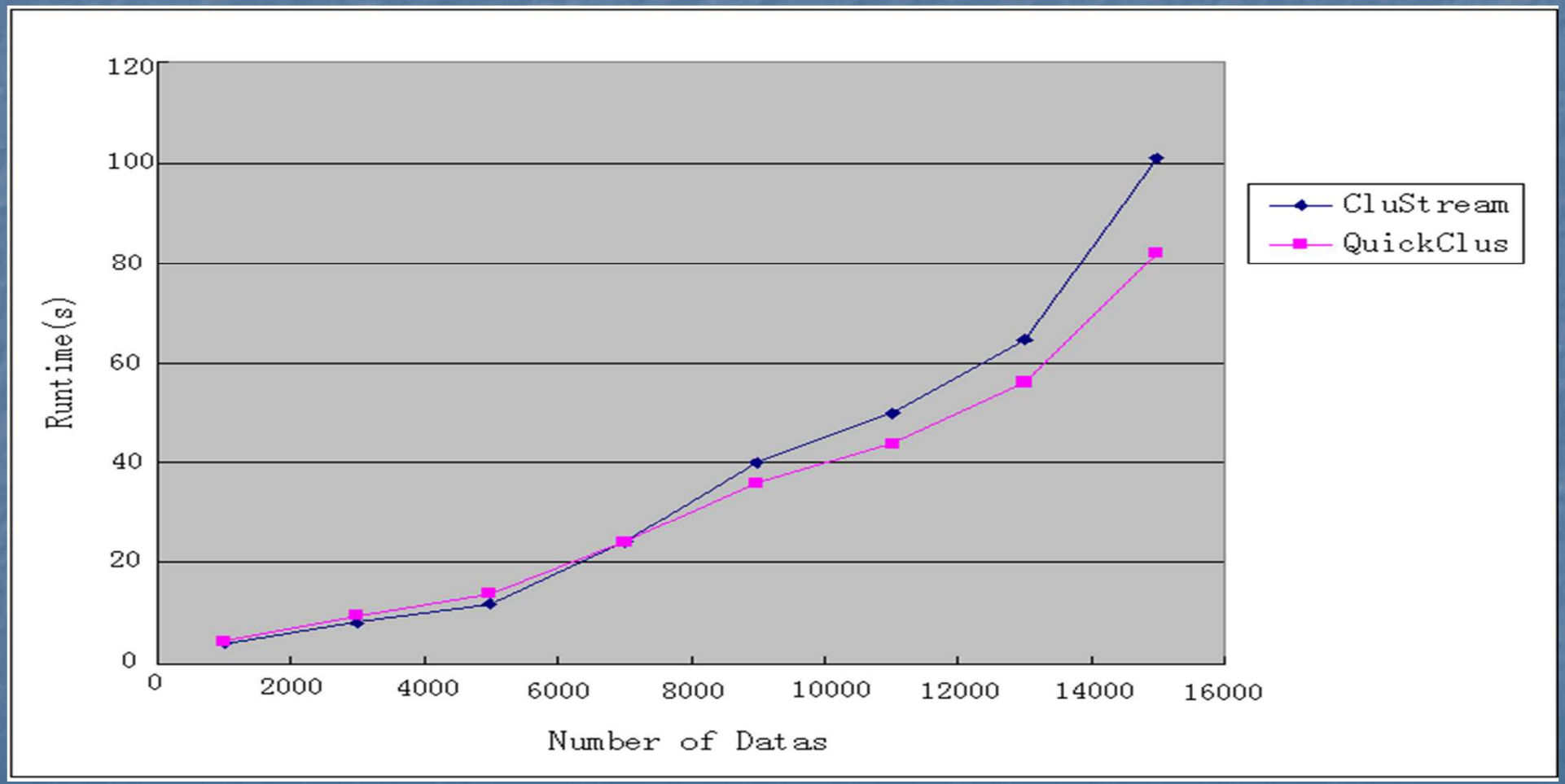
}

Просматривает каждый микро-кластер с порздком в hash-таблицу

{ if (плотность микро –кластера < D)

{if (значение плотность микро –кластер больше чем плотность выброса)
Резати этот микро-кластре
else
Опредить её единчной различной точкой , освобождение этот микро-кластер
}
else
{if (существует микро-кластер и значение плотности очень большой)
Сжатие этот микро-кластер
}
Сохранить данные по структуре Pyramid frame:
}
}

Сравнение времени кластеризации



Сравнение качества кластеризации