Методы кластерного анализа в задачах интеллектуальной обработки данных

доц. Нестеренко В.А.

Институт математики, механики и компьютерных наук

ЮФУ

Data Mining – обработка наборов данных с целью выявления ранее неизвестной, неочевидной, нетривиальной и практически полезной информации.

Имеется база данных (чем крупнее, тем лучше), в этой базе содержатся «скрытые знания», методы Data Mining позволяют выделить эти знания.

Методы Data Mining –

- ***кластерный анализ;***
- ***классификация данных;***
- ***нейронные сети;***
- ***математическая статистика;***
- ***генетические алгоритмы;***
- ***поиск явных и неявных ассоциаций;***
- ***визуализация данных;***
- ***. . .***

Кластеризация и классификация данных

- методы, основанные на мере схожести объектов, представленных в исходном наборе данных

Формализация метода:

- Набор объектов. Свойства или признаки объектов сводим к характеристикам.

- Вводим пространство характеристик, оси пространства соответствуют характеристикам. Объектам соответствуют точки в пространстве характеристик с соответствующими координатами.

- Далее, вводим метрику в пространстве характеристик – задаём расстояние между точками. Можем полагать, что чем ближе точки в пространстве характеристик, тем более похожи исходные объекты.

- ***Задача кластеризации: разбить исходное множество на группы (кластеры) так, чтобы объекты из одной группы были более похожи друг на друга чем объекты из разных групп.***

Характеристики или признаки объектов

Типы признаков:

- количественный: числовой
- порядковый: {ассистент, доцент, профессор}
- категориальный: {red, green, blue}
- бинарный: {0,1}, {true, false}, {male, female}

Мера близости (схожести) объектов

- Количественные признаки

  - расстояние между точками в пространстве характеристик (Евклидова метрика, …)

$$d^2\left(X^{(m)}, X^{(n)}\right) = \sum_i \left(x_i^{(m)} - x_i^{(n)}\right)^2$$

- Порядковые признаки

  - разность индексов в упорядоченной последовательности

$$d\left(X^{(m)}, X^{(n)}\right) = |m - n|$$

- Категориальные, бинарные признаки

  - число несовпадающих признаков для пары объектов (метрика Хэмминга)

$$d_H\left(X^{(m)}, X^{(n)}\right) = \sum_s \left(1 - \delta\left(x_s^{(m)}, x_s^{(n)}\right)\right)$$

Свойства кластеризации

Масштабная инвариантность:

Алгоритм кластеризации является ***масштабно-инвариантным***, если для любой функции расстояния ρ и любой константы α>0 результаты кластеризации с использованием расстояний ρ и α·ρ совпадают.

Свойства кластеризации

Согласованность:

Алгоритм кластеризации является ***согласованным***, если результат кластеризации не изменяется после допустимого преобразования функции расстояния ρ (расстояния между объектами внутри кластеров уменьшаются, вне кластеров - увеличиваются).

Свойства кластеризации

Полнота:

Алгоритм кластеризации является ***полным***, если множество результатов кластеризации алгоритма A при изменении функции расстояния ρ совпадает со множеством всех возможных разбиений множества объектов.

Пример: чётный - нечётный

Теорема Клейнберга о невозможности кластеризации

Для множества объектов, состоящего из двух и более элементов, не существует алгоритма кластеризации, который был бы одновременно масштабно-инвариантным, согласованным и полным.

(Kleinberg J. An Impossibility Theorem for Clustering)

НО можно построить алгоритм, удовлетворяющий любым двум аксиомам.

Метод К-средних (K-means)

Кластер $\boldsymbol{C_k}$ состоит из образующих его объектов X_i и описывается центром $X^{(k)}$ и радиусом $R^{(k)}$:

$$\boldsymbol{C_k} = \{X_i, i = 1,2, \dots \mid d(X^{(k)}, X_i) \leq R^{(k)}\}$$

$d(X^{(k)}, X_i)$ – расстояние от объекта до центра кластера.

Центр кластера

$$X^{(k)} = \frac{1}{N_k} \sum_{X_i \in \boldsymbol{C_k}} X_i$$

- геометрическое среднее образующих его точек

Радиус кластера

$$R^{(k)} = \frac{A}{N_k} \sum_{X_i \in \boldsymbol{C_k}} \left(X_i - X^{(k)}\right)^2$$

- определяется среднеквадратичным отклонение точек от центра

Идея метода, минимизировать суммарное расстояние точек от центров соответствующих кластеров:

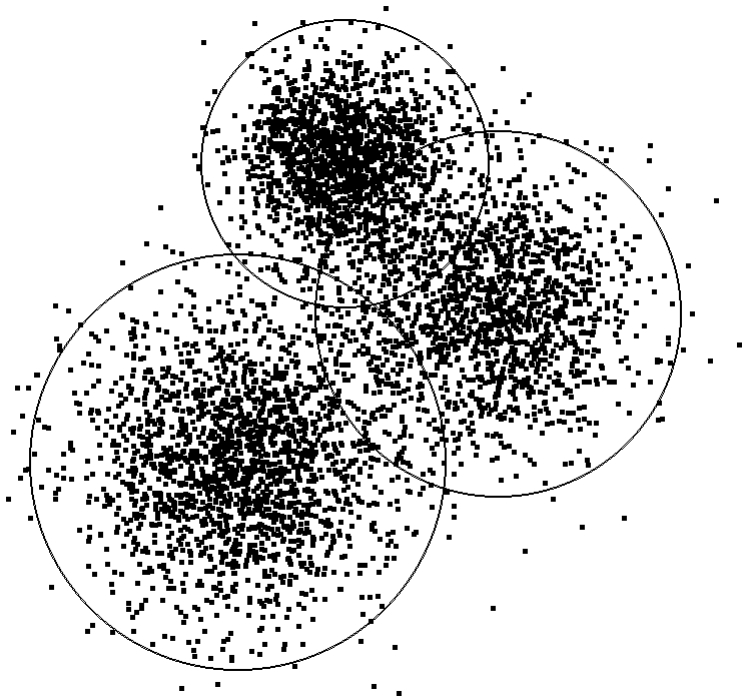
$$\min \sum_{\boldsymbol{C_k}} \sum_{\boldsymbol{X_i \in C_k}} \boldsymbol{d(X^{(k)}, X_i)}$$

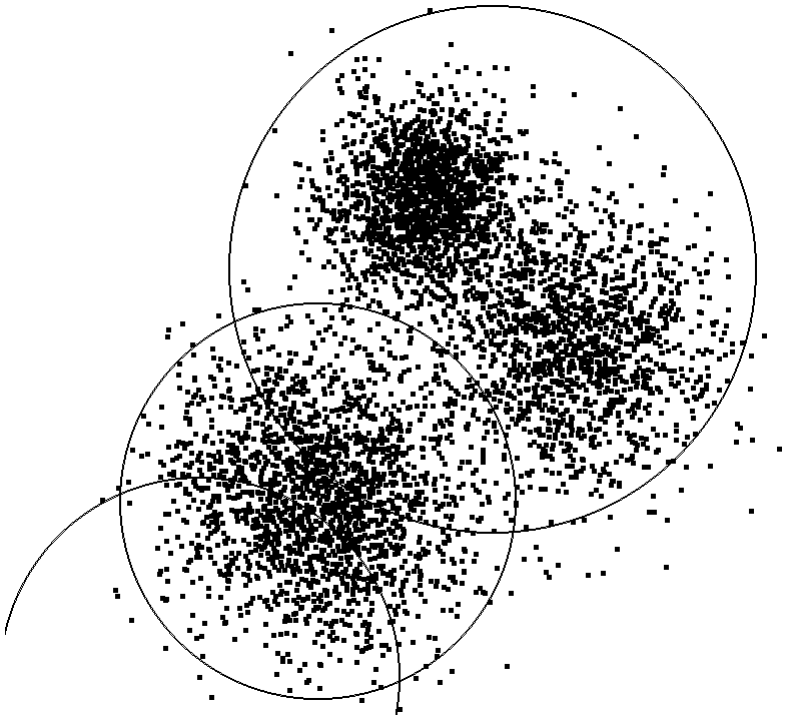
Метод K-средних (K-means)

Алгоритм:

1. Начинаем с пустого набора кластеров **Ω** и затравочного радиуса R_0
2. Для точки X_i исходного множества ищем ближайший кластер $\boldsymbol{C}_k$, внутрь которого попадает данная точка.
3. Если такой кластер найден, то включаем точку X_i в состав кластера $\boldsymbol{C}_k$ и пересчитываем центр $X^{(k)}$ и радиус $R^{(k)}$ кластера $\boldsymbol{C}_k$ с учётом новой точки.
4. Если подходящий кластер не найден, то создаём новый кластер с центром в точке X_i и затравочным радиусом R_0.
5. Если два кластера оказываются достаточно близко, то эти два кластера объединяются в один.
6. Возвращаемся к шагу 2 для обработки следующей точки.

Метод K-средних (пример):





Метод нечёткой кластеризации (C-means, Fuzzy clustering)

Этот метод похож на метод K-means. Вхождение точки в кластер определяется на основе матрицы принадлежности: r_{ij} - вероятность вхождения элемента X_i в кластер $\boldsymbol{C}_j$. Элементы матрицы принадлежности зависят от степени близости точки к центру кластера.

Например, в предположении о нормальном законе распределения относительно центра кластера:

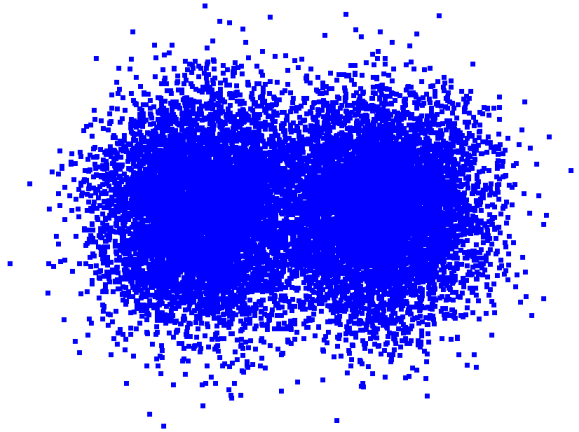
$$r_{ij} = \frac{\mathcal{N}\left(d\left(X^{(j)}, X_i\right) \middle| \mu = 0, \sigma^{(j)}\right)}{\sum_k \mathcal{N}\left(d\left(X^{(k)}, X_i\right) \middle| \mu = 0, \sigma^{(k)}\right)}$$

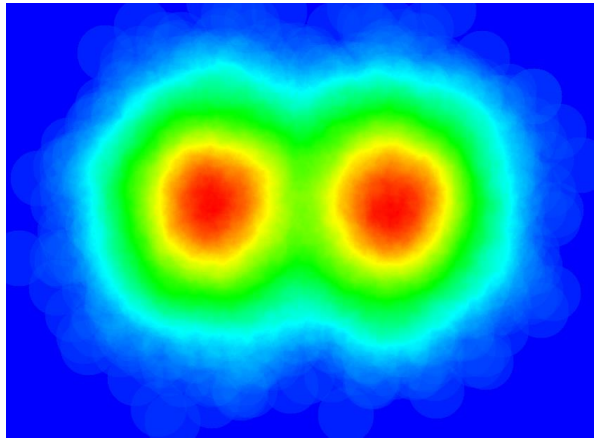
$\mathcal{N}$ - плотность вероятности нормального распределения.

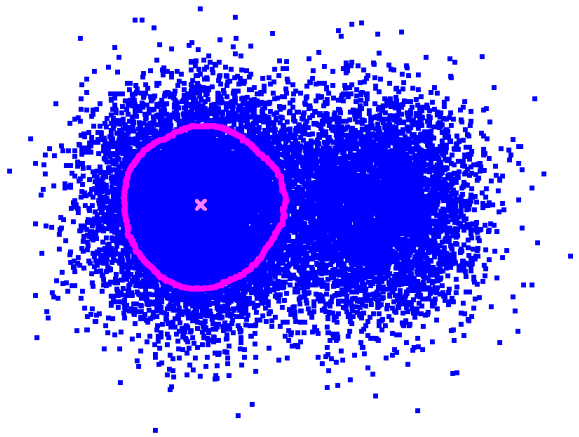
В остальном аналогично методу K-средних за исключением того, что требуется изначально задавать число кластеров и, возможно, их начальное положение.

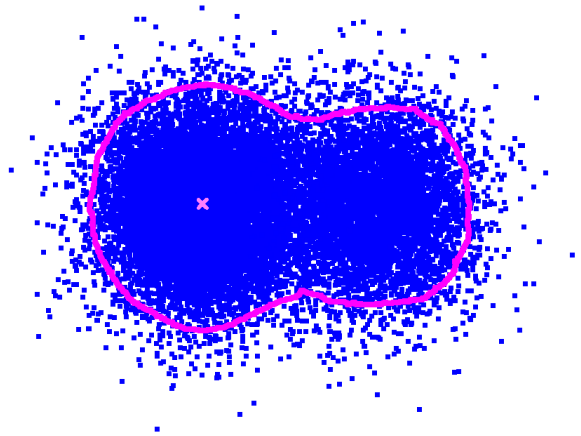
Недостатки: Центроидные кластеры. Требуется задавать число кластеров и их начальное положение.

Метод кластеризации на основе плотности









Другие методы кластеризации

1. *Метод ближайших соседей.*
2. *Иерархическая кластеризация.*
3. *. . .*

Критерии качества кластеризации

Задача кластеризации в терминах множества точек в пространстве характеристик: разбить исходное множество на группы (кластеры) так, чтобы точки из одного кластеры располагались ближе друг к другу чем точки из разных кластеров.

Это определение даёт вполне очевидный и интуитивно понятный критерий оценки качества кластеризации: пусть

$$D_k = \sum_{X_i \in \boldsymbol{C}_k} d^2 \left(X_i - X^{(k)}\right)$$

тогда критерий

$$K = \frac{D_{ab} - (D_a + D_b)}{D_a + D_b}$$

позволяет оценить результат слияния кластеров $\boldsymbol{C}_a$ и $\boldsymbol{C}_b$: $\boldsymbol{C}_{ab} = \boldsymbol{C}_a \cup \boldsymbol{C}_b$

(В частности, позволяет выбрать число кластеров)

Пространство характеристик, размерность

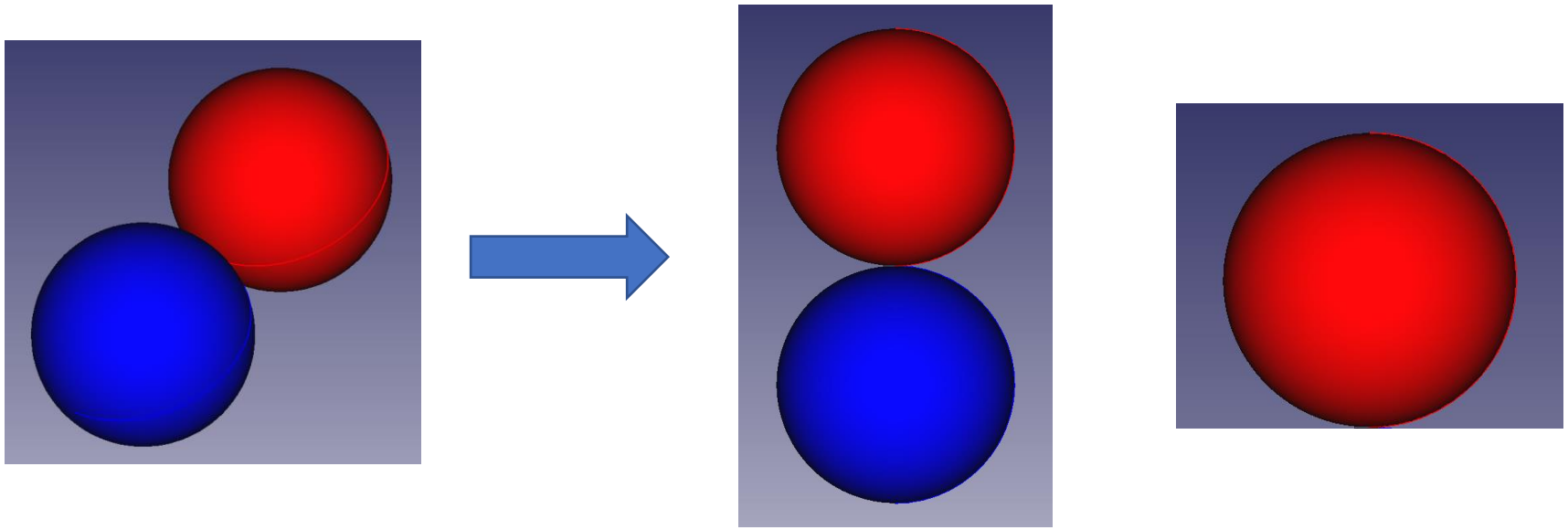
Объекты -> свойства объектов, характеристики -> пространство характеристик -> точки в пространстве характеристик -> метрика

Схожесть объектов ↔ расстояние между точками

1. Метод главных компонент.
2. Использование стресс-функции.

Пространство характеристик, размерность

Снижение размерности пространства характеристик – проецирование в пространство меньшей размерности (по возможности) без потери информации.



Редукция размерности, метод главных компонент

Выбор проекционного подпространства определяется направлениями наибольшего разброса точек в пространстве характеристик

Математически это сводится к решению задачи на собственные значения для ковариационной матрицы:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}$$

$$\sigma_{kl} = \frac{1}{N} \sum_{X_i} X_i^k X_i^l$$

k, l – номер характеристики.

Редукция размерности, стресс - функция

Выбор проекционного подпространства определяется из условия минимума потери информативности данных

D_{ij}^2 - квадрат расстояния между объектами X_i и X_j в исходном пространстве характеристик

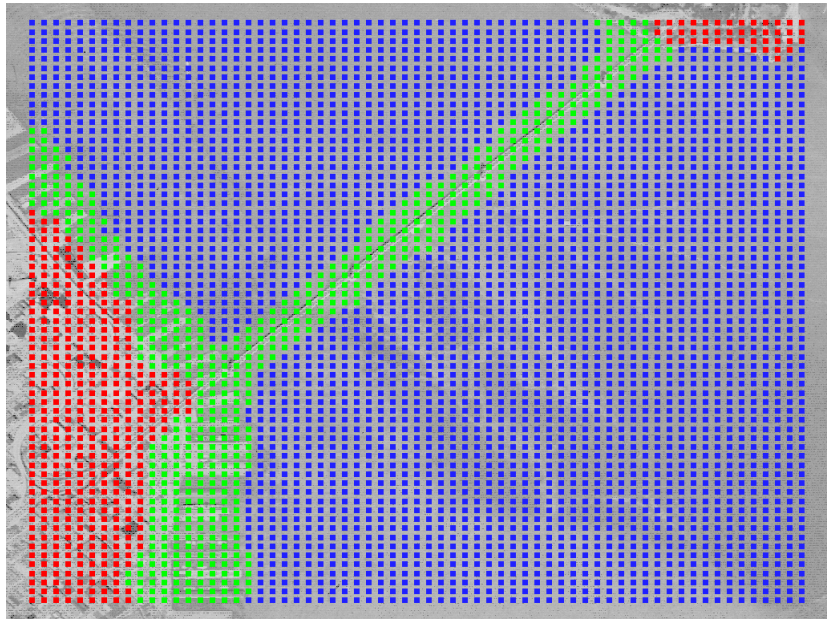
d_{ij}^2 - квадрат расстояния между объектами в редуцированном пространстве

Проекция в редуцированное пространство выбирается так, что бы минимизировать изменение расстояний между объектами:

$$F_s = \sum_{i,j} (D_{ij}^2 - d_{ij}^2)^2$$

Сегментация изображений

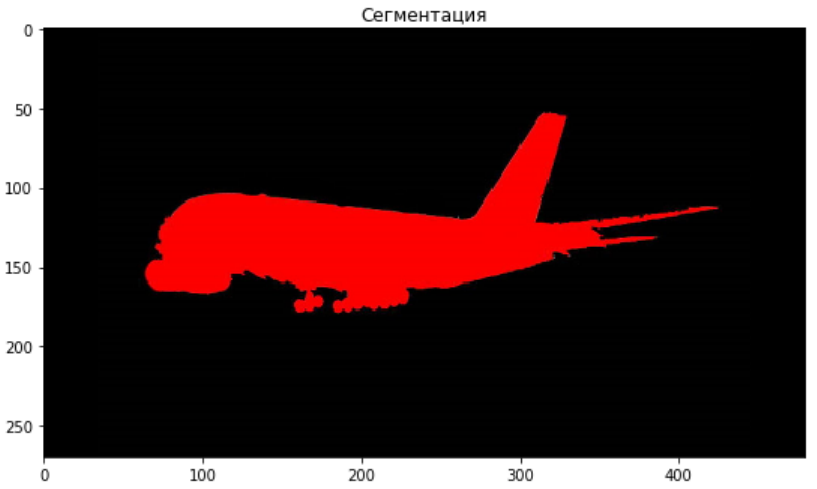
Сегментация изображений — это процесс разбиения пикселей изображения на группы так, чтобы пиксели со сходными визуальными характеристиками попали в одинаковые группы.



Сегментация изображений

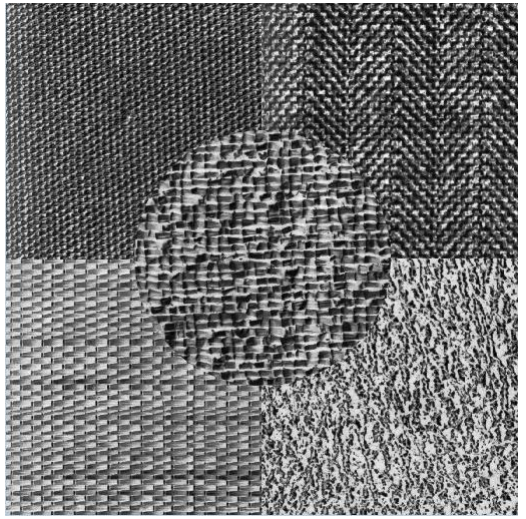
Признаки и характеристики пикселей:

1. *Цвет*
2. *Яркость*
3. *. . .*

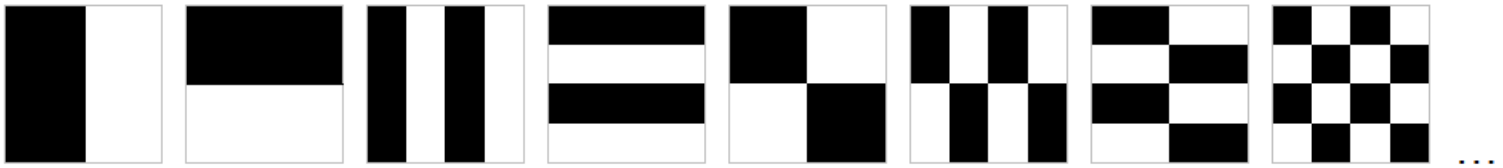


Сегментация изображений

1. ~~Цвет~~
2. ~~Яркость~~
3. *Координаты в изображении*
4. *Соседние пиксели*
5. *. . .*

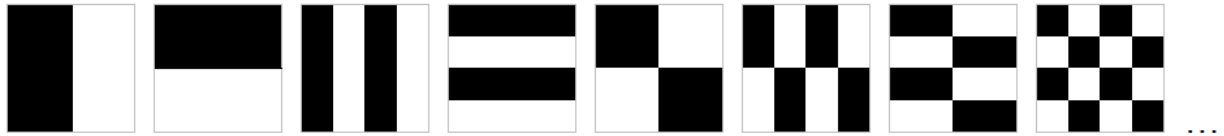


Маски Хаара:



Сегментация изображений

Маски Хаара:

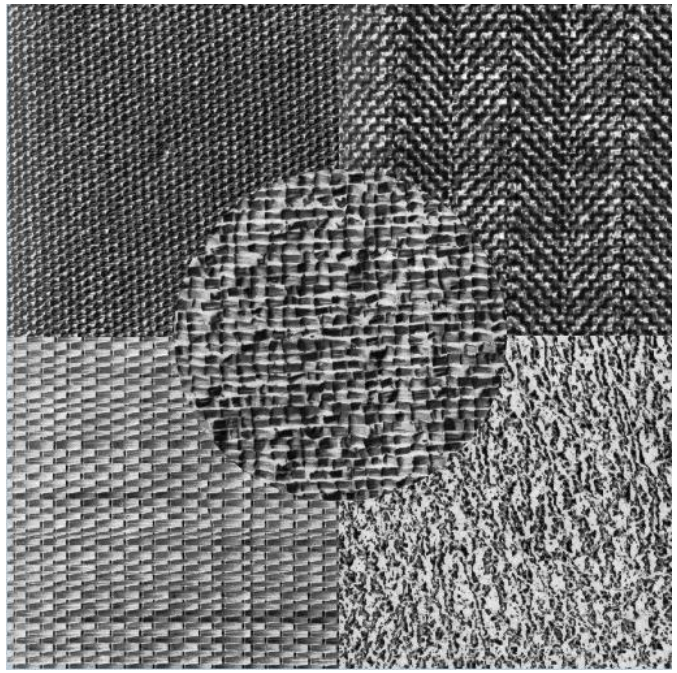


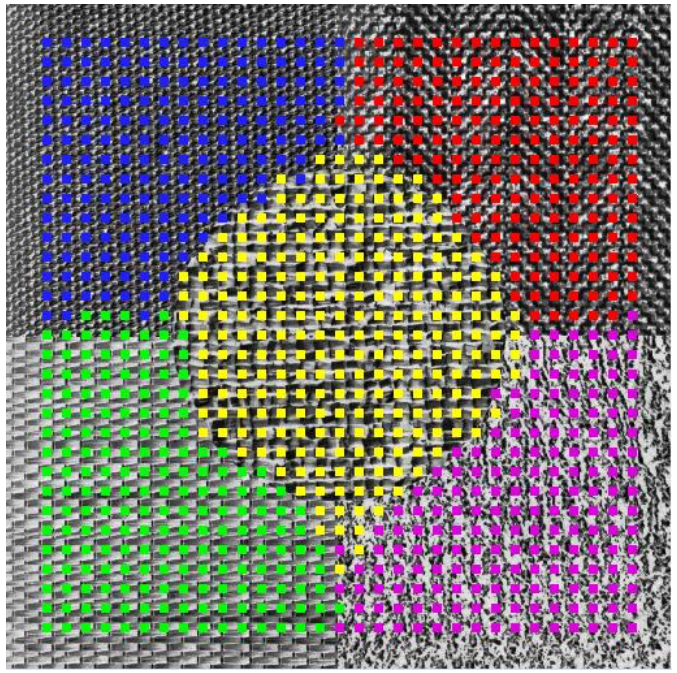
Размер: 60x60 пикселей

Число разбиений: {1; 2; 3; 4; 5; 6; 10; 12; 15; 20; 30}

Исключаем: {1; 3; 5 15} x {1; 3; 5 15}

В результате 105 характеристик

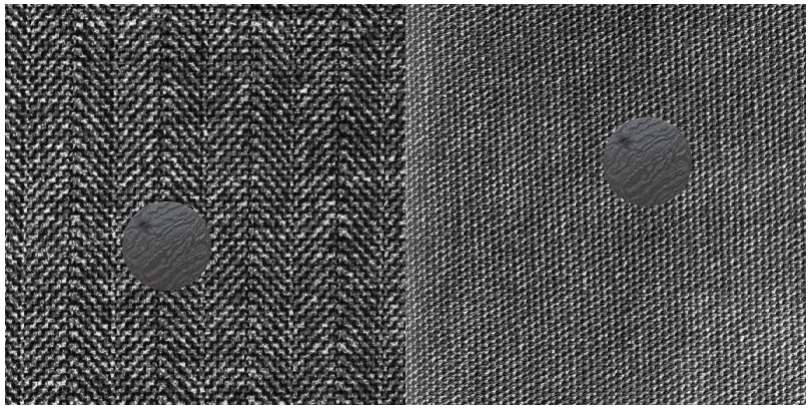


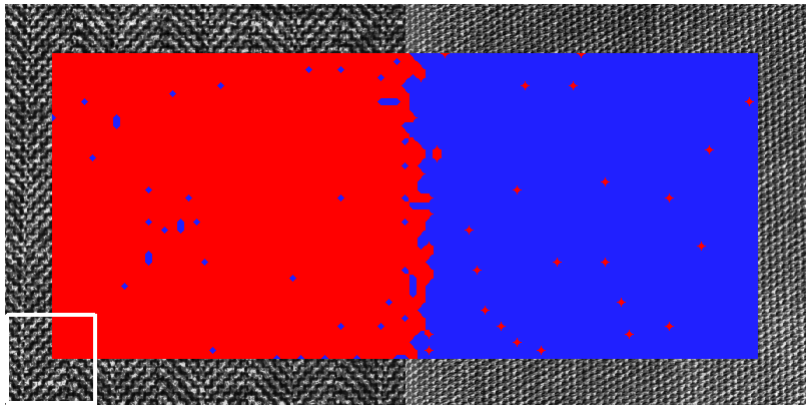


Сегментация изображений

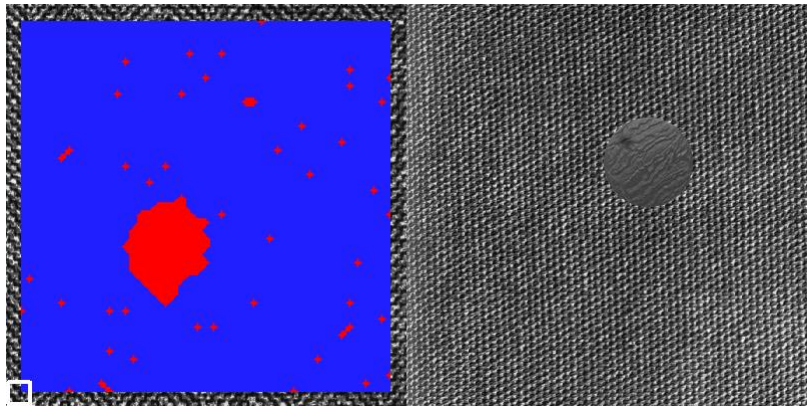
При изменении размера маски меняется детализация.

Размер маски 90x90



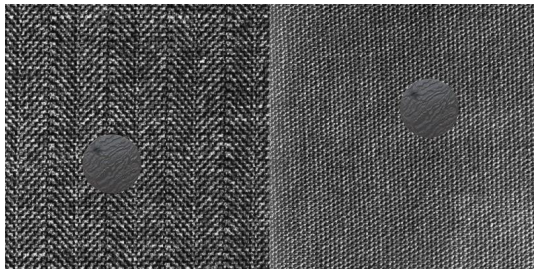


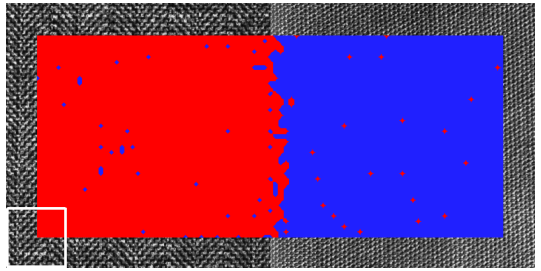
Размер маски 24x24



Сегментация изображений

Число кластеров (сегментов)





Критерий

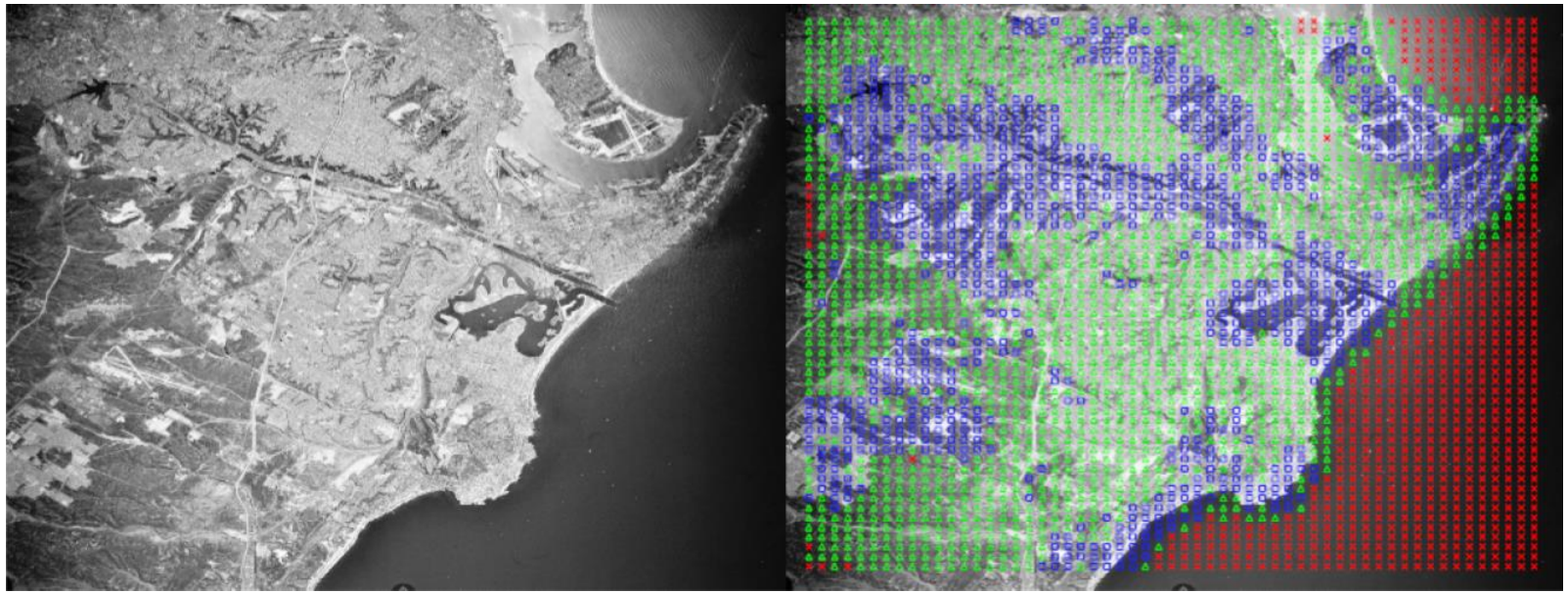
$$K = \frac{D_{ab} - (D_a + D_b)}{D_a + D_b}$$

позволяет оценить результат слияния кластеров $\boldsymbol{C}_a$ и $\boldsymbol{C}_b$: $\boldsymbol{C}_{ab} = \boldsymbol{C}_a \cup \boldsymbol{C}_b$

Число кластеров	Критерий К
5	0.003
4	0.009
3	0.055
2	0.156

Сегментация изображений

Цель сегментации заключается в упрощении изображения, чтобы его было проще и легче анализировать.



Data Mining в задачах обнаружения нарушений информационной безопасности

Выявление нарушений. Два пути:

1. Мы знаем в чём заключается нарушение, известны его признаки. По этим признакам мы выявляем нарушение (обучение с учителем).
2. Мы не знаем признаков нарушения (новый тип), нам известны признаки отсутствия нарушения – нормальное состояние системы. Отклонения от нормального состояния (аномалии) являются сигналом о возможной атаке на систему (обучение без учителя).

Дороти Деннинг (D. E. Denning, 1987) - использование уязвимостей системы проявляется в аномальном (нетипичном) поведении системы. Поэтому, нарушения безопасности могут быть обнаружены при появлении признаков неправильного поведения системы.

Data Mining. Выявление аномалий

- Система, безопасность которой нас интересует.
- Объект наблюдения – характеризует состояние системы.
- Характеристики объекта наблюдения.
- Набор множества наблюдений –> множество точек в пространстве характеристик.
- Применение методов Data Mining

Адекватный выбор характеристик приведёт к разделению нормальных и аномальных состояний в пространстве характеристик.

Большая часть наблюдений соответствует нормальному состоянию системы (не всё безнадёжно потеряно).

Data Mining. Безопасность сети (пример)

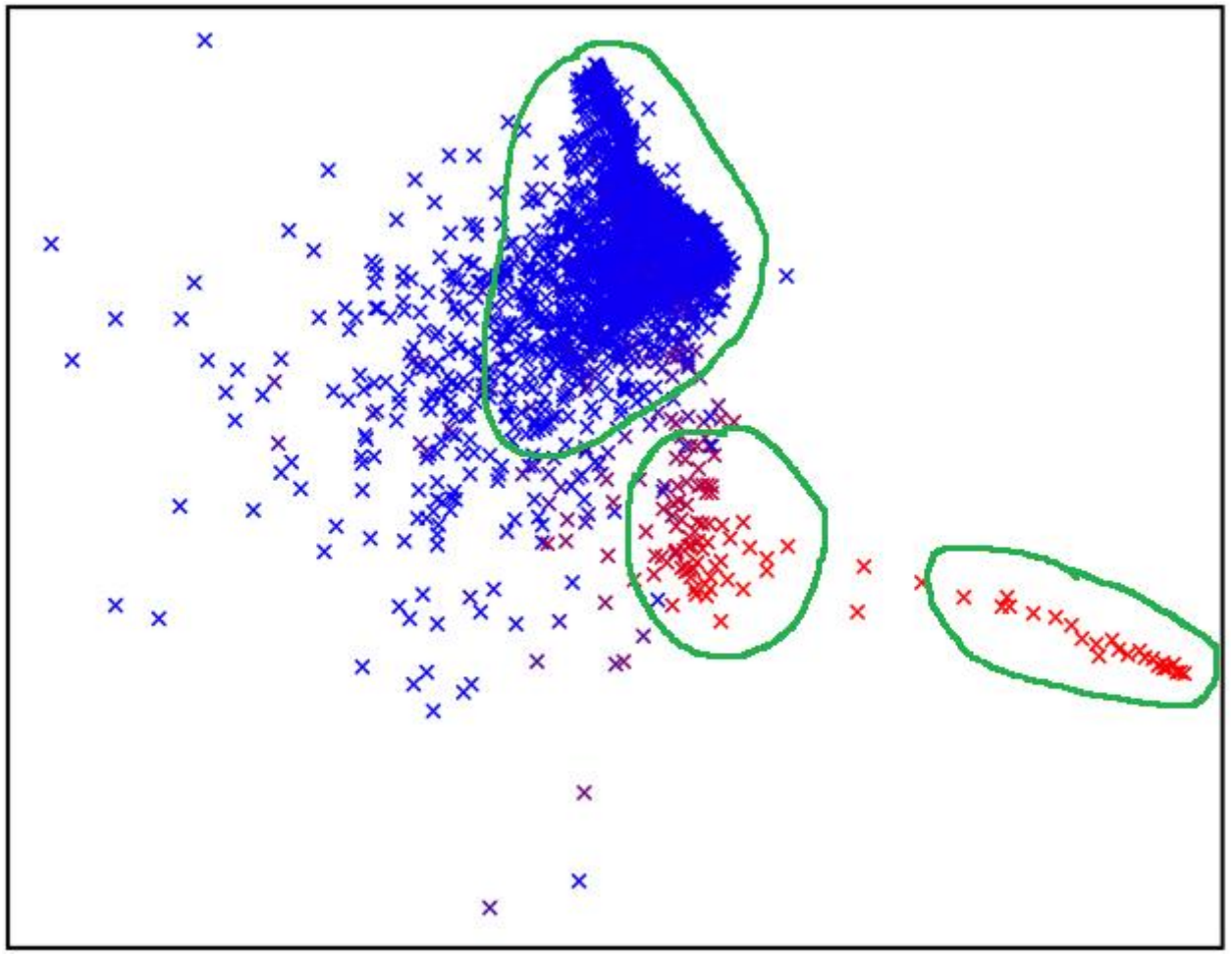
Объект наблюдения: набор пакетов за определённый интервал времени (временное окно 1-10 сек).

Характеристики объекта (в окне):

1. Временной интервал между двумя последними пакетами с одного узла.
2. Наибольшее значение числа пакетов с одного узла.
3. Среднее значение разницы идентификаторов двух последних пакетов.
4. Среднее значение размера пакета.
5. Разница между числом syn и synack -пакетов.

Data Mining. Безопасность сети (пример)

Реальный пример:



Data Mining. Достоверность результатов

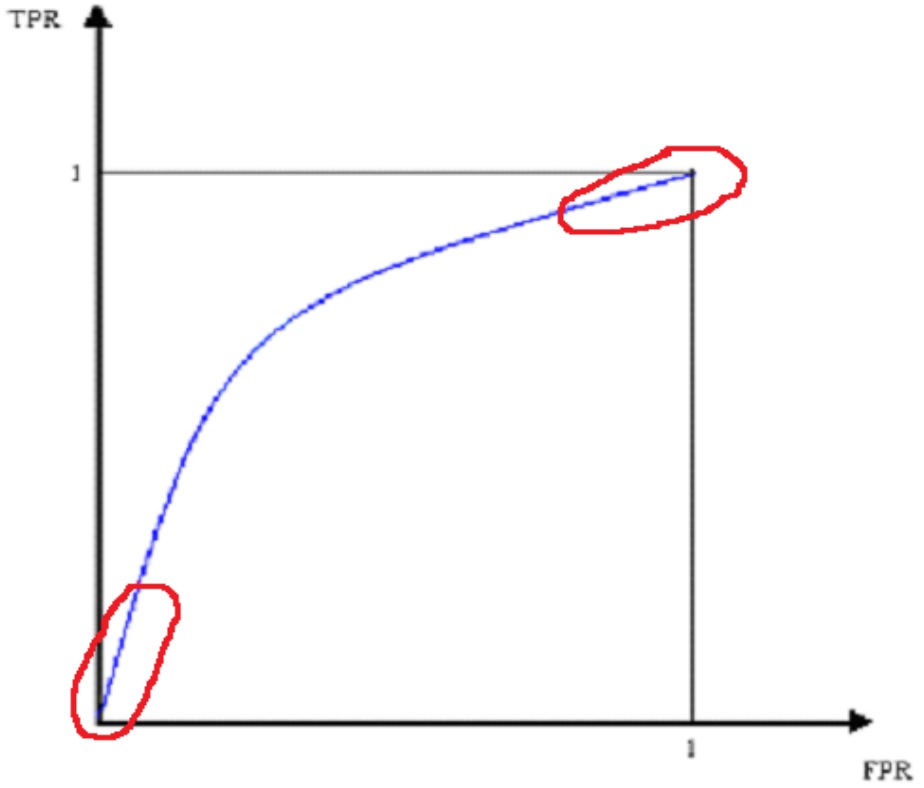
$$\text{FPR (False Positive Rate)} = \frac{\text{число нормальных событий отмеченных как атака}}{\text{общее число нормальных событий}}$$

$$\text{TPR (True Positive Rate)} = \frac{\text{число правильно атака}}{\text{общее число атак}}$$

Идеальная система:

FPR = 0

TPR = 1



Источники:

- Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP | Холод Иван Иванович, Степаненко Валентин Владимирович

- Заглавная страница (machinelearning.ru)
- Bishop_-_Pattern_Recognition_and_Machine_Learning(proglib).pdf (vk.com)
- Data_Mining_-_Concepts_and_Techniques_(3rd_Ed)(proglib).pdf (vk.com)
- Data_Mining_Practical_Machine_Learning_Tools_and_Techniques(proglib).pdf (vk.com)
- Data Mining Course (kdnuggets.com)
- Kaggle: Your Home for Data Science