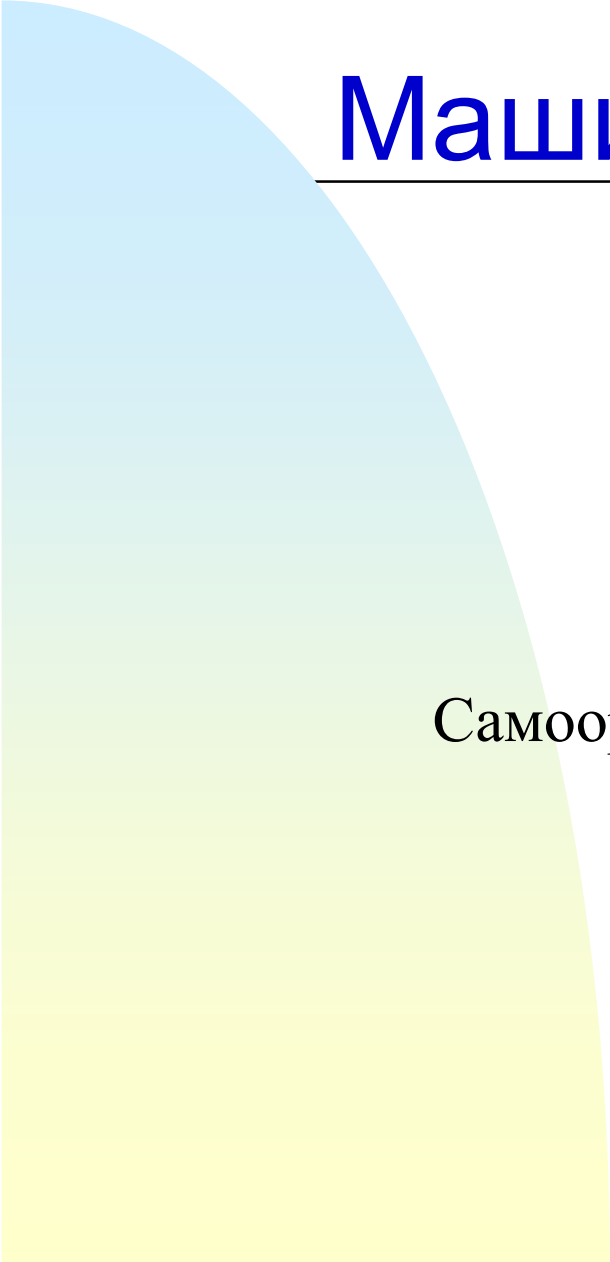




Машинное обучение

Лекция 3. Часть 1.
Самоорганизующиеся карты Кохонена

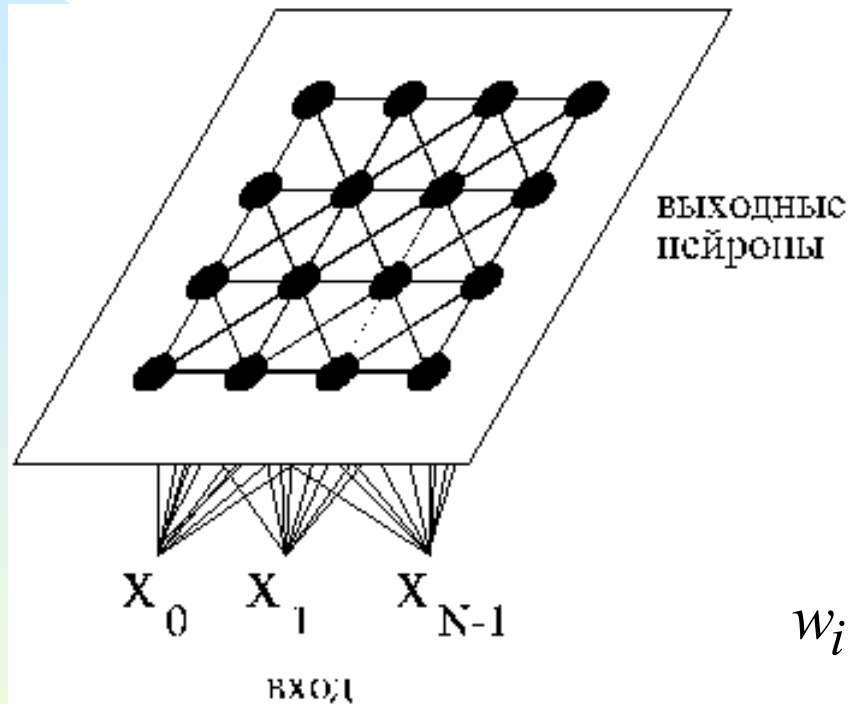
- Идея сети Кохонена принадлежит финскому ученому Тойво Кохонену (1982 год).
- Основной принцип работы сетей - введение в правило обучения нейрона информации относительно его расположения.
- В основе идеи сети Кохонена лежит аналогия со свойствами человеческого мозга. Кора головного мозга человека представляет собой плоский лист и свернута складками. Таким образом, можно сказать, что она обладает определенными топологическими свойствами (участки, ответственные за близкие части тела, примыкают друг к другу и все изображение человеческого тела отображается на эту двумерную поверхность).
- Сетью Кохонена реализуется классификация без учителя

- Сеть Кохонена обучается методом последовательных приближений. В процессе обучения таких сетей на входы подаются данные, но сеть при этом подстраивается не под эталонное значение выхода, а под закономерности во входных данных. Начинается обучение с выбранного случайным образом выходного расположения центров.
- В процессе последовательной подачи на вход сети обучающих примеров определяется наиболее схожий нейрон (тот, у которого скалярное произведение весов и поданного на вход вектора минимально). Этот нейрон объявляется победителем и является центром при подстройке весов у соседних нейронов. Такое правило обучения предполагает "соревновательное" обучение с учетом расстояния нейронов от "нейрона-победителя".
- Обучение при этом заключается не в минимизации ошибки, а в подстройке весов (внутренних параметров нейронной сети) для наибольшего совпадения с входными данными.

- Основной итерационный алгоритм Кохонена последовательно проходит ряд эпох, на каждой из которых обрабатывается один пример из обучающей выборки. Входные сигналы последовательно предъявляются сети, при этом желаемые выходные сигналы не определяются. После предъявления достаточного числа входных векторов синаптические веса сети становятся способны определить кластеры. Веса организуются так, что топологически близкие узлы чувствительны к похожим входным сигналам.
- В результате работы алгоритма центр кластера устанавливается в определенной позиции, удовлетворительным образом кластеризующей примеры, для которых данный нейрон является "победителем". В результате обучения сети необходимо определить меру соседства нейронов, т.е. окрестность нейрона-победителя.

- Окрестность представляет собой несколько нейронов, которые окружают нейрон-победитель
- Сначала к окрестности принадлежит большое число нейронов, далее ее размер постепенно уменьшается. Сеть формирует топологическую структуру, в которой похожие примеры образуют группы примеров, близко находящиеся на топологической карте.
- Полученную карту можно использовать как средство визуализации при анализе данных. В результате обучения карта Кохонена классифицирует входные примеры на кластеры (группы схожих примеров) и визуально отображает многомерные входные данные на плоскости нейронов.
- Уникальность метода самоорганизующихся карт состоит в преобразовании n -мерного пространства в двухмерное. Применение двухмерных сеток связано с тем, что существует проблема отображения пространственных структур большей размерности.
- Имея такое представление данных, можно визуально определить наличие или отсутствие взаимосвязи во входных данных.
- Нейроны карты Кохонена располагают в виде двухмерной матрицы, раскрашивают эту матрицу в зависимости от анализируемых параметров нейронов.

Сеть с самоорганизацией на основе конкуренции



$$w_i = (w_{i0}, w_{i1}, \dots, w_{i(N-1)})^T$$

$$d(x, w_j) = \min_{0 \leq i < N} d(x, w_i)$$

j-ый нейрон - победитель

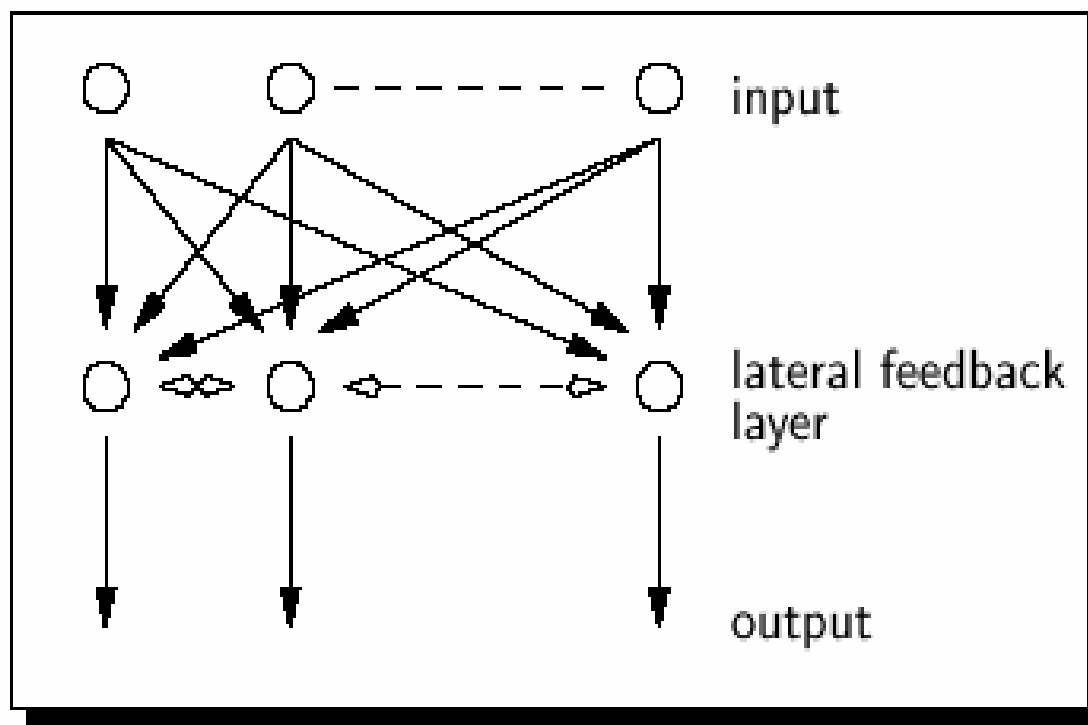
$$w_i(k+1) = w_i(k) + \eta_i(k)[x - w_i(k)]$$

Модификация весов по правилу Кохонена.

Изменяются только веса, лежащие в некоторой окрестности нейрона-победителя.

Окрестность задается функцией близости – может быть пороговой или Гауссовой

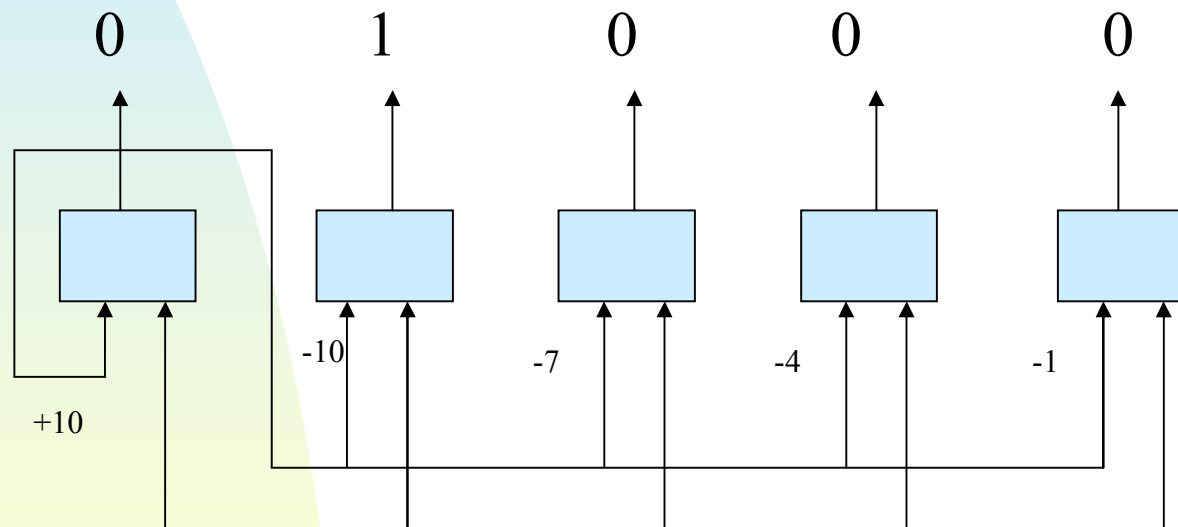
Структура сети Кохонена



Латеральные (тормозящие) связи обеспечивают срабатывание (победу) одного нейрона.

Пример латеральных связей

Показаны обратные (латеральные) связи только для одного нейрона (показаны их веса)



Входной вектор

Меры расстояния между векторами

евклидова мера: $d(x, w_i) = \|x - w_i\| = \sqrt{\sum_j (x_j - w_{ij})^2}$

скалярное произведение: $d(x, w_i) = 1 - \|x\| \|w_i\| \cos(x, w_i)$

Манхэттенское
расстояние:

$$d(x, w_i) = \sqrt{\sum_j |x_j - w_{ij}|}$$

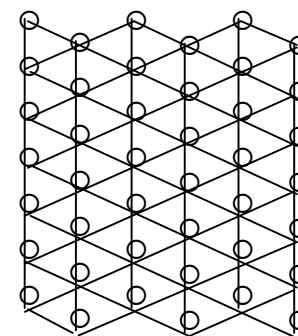
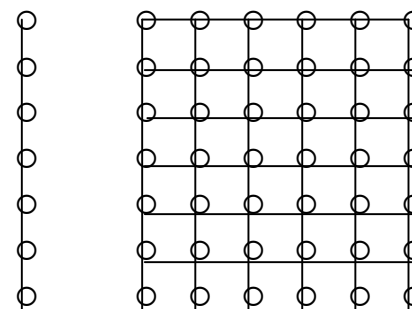
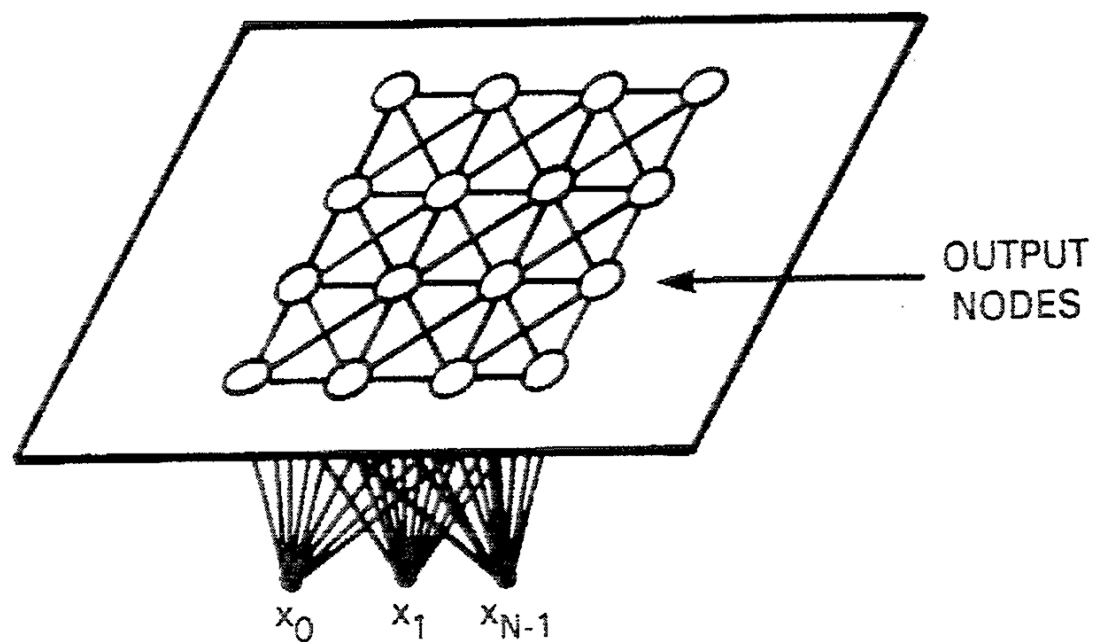
мера относительно нормы L_∞ : $d(x, w_i) = \max_j |x_j - w_{ij}|$

Где:

x — входной вектор

w_i — вектор весов i -го нейрона

Шаг1. Выбор топологии



Шаг 2. Нормализация векторов

Необходимость в нормализации при небольшой размерности входных векторов.

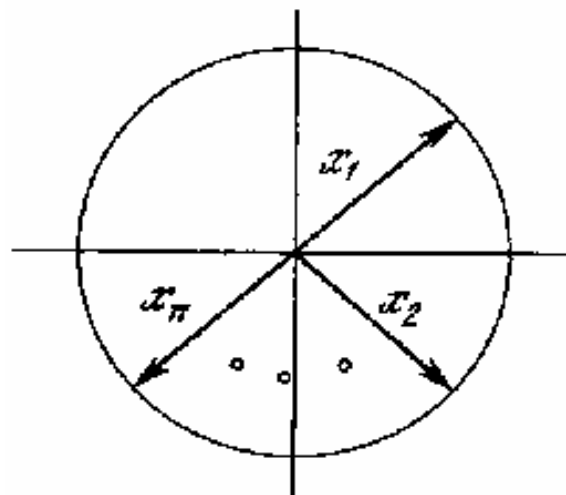
При $N > 200$ эффект от нормализации уменьшается.

Способы нормализации:

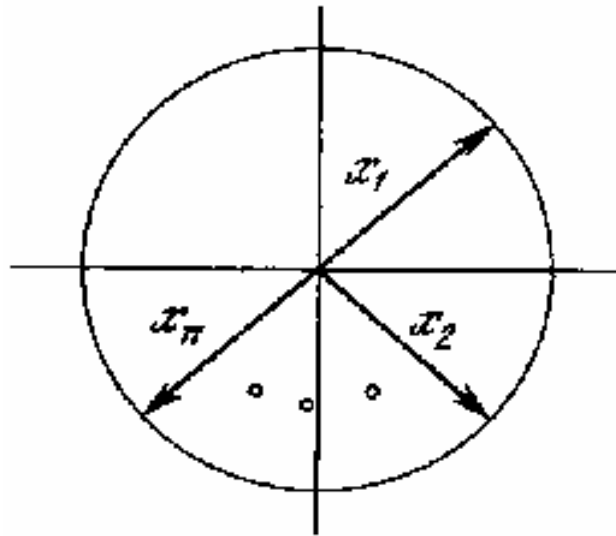
1.
$$x_i \leftarrow \frac{x_i}{\sqrt{\sum_i x_i^2}}$$

2. Увеличение размерности пространства на одну координату, чтобы:

$$\sum_{i=0}^N x_i^2 = 1$$



Шаг 3. Инициализация весов



1. Инициализация случайным образом (с нормализацией).
2. Метод выпуклой комбинации:

$$w_{ij} = \frac{1}{\sqrt{N}}$$

$$x_i = \alpha x_i + \frac{1-\alpha}{\sqrt{N}}$$

3. Добавление шума к входным векторам.

Шаг 4. Обучение. Алгоритмы обучения

$$E = \frac{1}{N} \sum_{i=1}^N \|x_i - w_{w(i)}\|^2$$

1. WTA с учетом активности нейронов (CWTA):

$$w_h = w_h + \eta[x - w_h]$$

2. WTM (Winner takes most):

$$w_i = w_i + \eta_i G(i, x)[x - w_i]$$

а) классический алгоритм Кохонена:

- функция G определяет прямоугольную окрестность;
- соседство гауссовского типа:

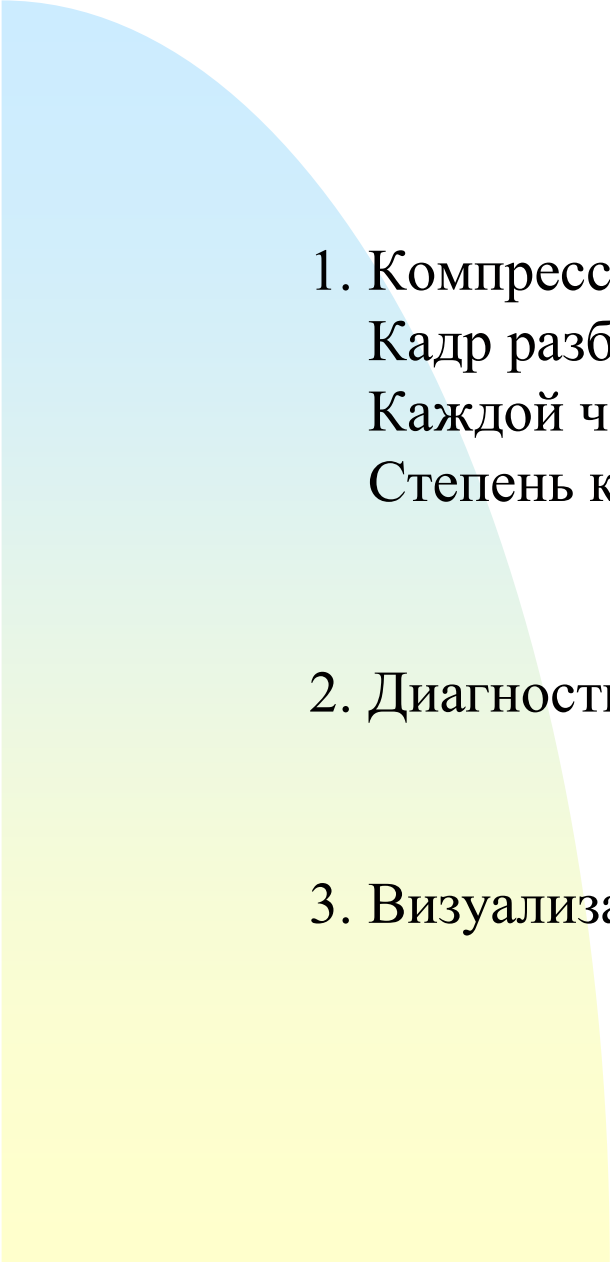
$$G(i, x) = e^{-\frac{d^2(i, w)}{2\lambda^2}}$$

б) алгоритм нейронного газа:

Сортировка векторов весов по расстояниям до входного вектора;

$G(i, x) = e^{-\frac{m(i)}{\lambda}}$, где $m(i)$ - номер нейрона i в полученной перестановке по расстояниям.

Области применения сети Кохонена



1. Компрессия данных.

Кадр разбивается на части (входные вектора).

Каждой части сопоставляется нейрон-победитель.

Степень компресии изображений: 16.

2. Диагностирование неисправностей.

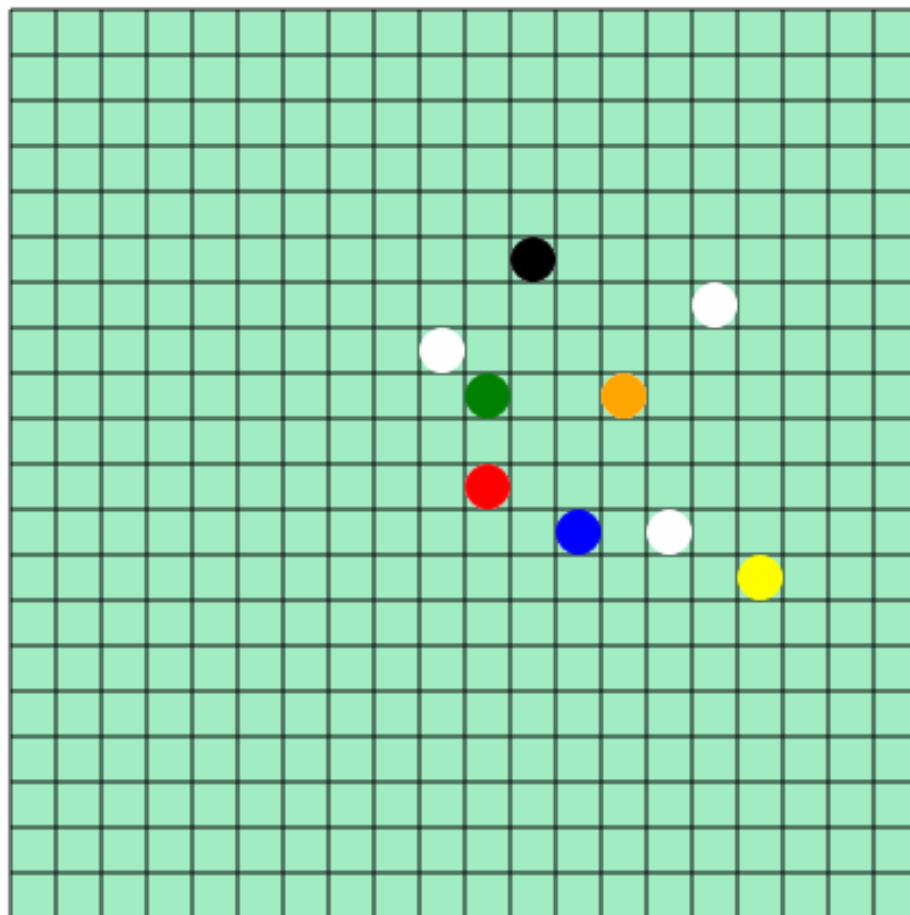
3. Визуализация многомерной информации.

Классификация танков 2-ой мировой войны (Из лекций Шамина)

Название	Масса	Длина	Ширина	Мощность двигателя	Скорость
PzKpfw I	5.4	4020	2060	57	37
PzKpfw II	8.9	4810	2223	140	40
PzKpfw III	19.5	5380	2910	285	60
PzKpfw IV	25	5890	2880	300	40
«Пантера»	44.8	6870	3270	700	55
«Тигр»	56	6316	3705	700	44
«Королевский тигр»	68	7380	3755	700	38
T-50	13.8	5200	2470	300	60
T-70	9.2	4285	2348	140	42
T-34	30	5920	3000	500	54
T-44	31.8	6070	3180	500	60
KB-1	47.5	6675	3320	600	34
ИС-1	44.2	6770	3070	520	37
ИС-2	46	6770	3070	520	37
ИС-3	49	6900	3150	520	40

Результат

Self-Organizing Map



↔ — □ ×

Построить карту

Показать данные

Проверить

- ☒ Black
- ☐ Blue
- ☐ Red
- ☐ Green
- ☐ Yellow
- ☐ Orange

Выход

«Тигр»

T-34

PzKpfw III

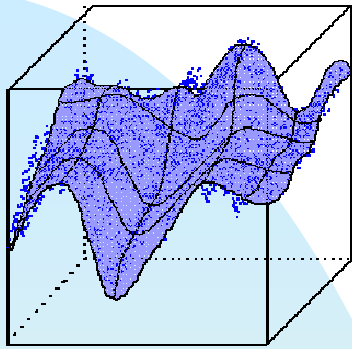
«Пантера»

PzKpfw I

ИС-3

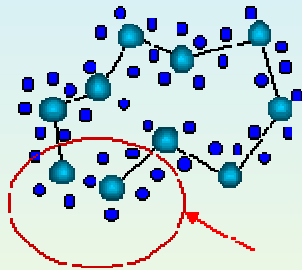
Это результат градуирования – наложения примеров на карту с привязкой примера к ближайшему нейрону

Визуализация многомерной информации

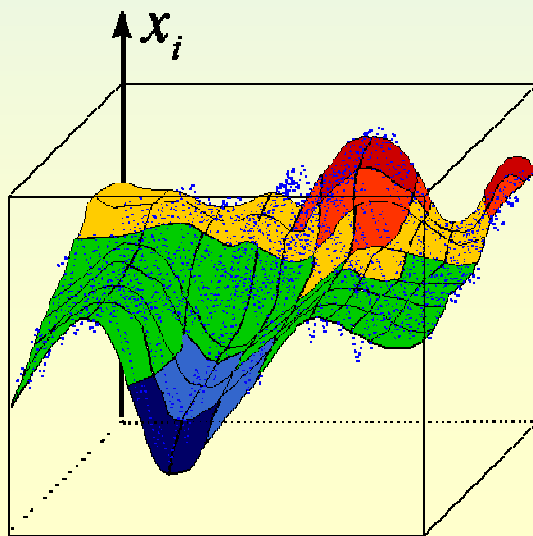


Двумерная топографическая карта набора трехмерных данных.

Каждая точка в трехмерном пространстве попадает в свою ячейку сетки имеющую координату ближайшего к ней нейрона из двумерной карты.



Свойство локальной близости: близкие на карте области близки в исходном пространстве. Наоборот в общем случае не верно.

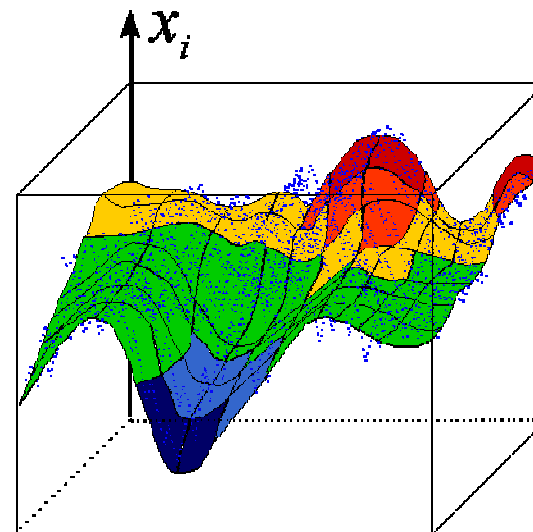
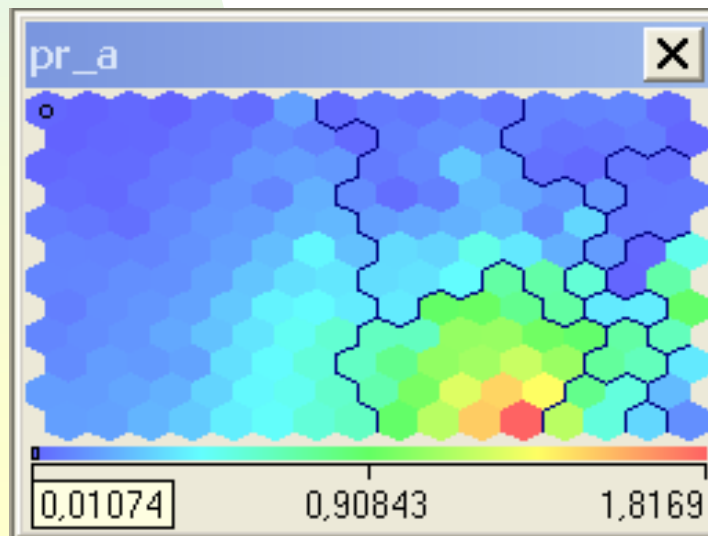


Раскраска по i -му признаку.

Карты всех признаков образуют топографический атлас.

Раскраска карт Кохонена

- Раскраска карты (ее i -го признака – pr_a) в трехмерном представлении.
- Темно-синие участки на карте соответствуют наименьшим значениям показателя, красные - самым высоким.



- Таким образом, карты Кохонена (как и географические карты) можно отображать:
 - ◆ в двухмерном виде, тогда карта раскрашивается в соответствии с уровнем выхода нейрона;
 - ◆ в трехмерном виде.
- В результате работы алгоритма получаем такие карты:
 - ◆ карта входов нейронов ;
 - ◆ карта выходов нейронов ;
 - ◆ специальные карты.
- Координаты каждой карты определяют положение одного нейрона. Так, координаты [15:30] определяют нейрон, который находится на пересечении 15-го столбца с 30-м рядом в матрице нейронов.

Карта входов нейронов

- Веса нейронов подстраиваются под значения входных переменных и отображают их внутреннюю структуру. Для каждого входа рисуется своя карта, раскрашенная в соответствии со значением конкретного веса нейрона.
- При анализе данных используют несколько карт входов.
- На одной из карт выделяют область определенного цвета - это означает, что соответствующие входные примеры имеют приблизительно одинаковое значение соответствующего входа. Цветовое распределение нейронов из этой области анализируется на других картах для определения схожих или отличительных характеристик.

Карта выходов нейронов

- На карту выходов нейронов проецируется взаимное расположение исследуемых входных данных. Нейроны с одинаковыми значениями выходов образуют кластеры - замкнутые области на карте, которые включают нейроны с одинаковыми значениями выходов.

Специальные карты

- Это карта кластеров, матрица расстояний, матрица плотности попадания и другие карты, которые характеризуют кластеры, полученные в результате обучения сети Кохонена.
- Важно понимать, что между всеми рассмотренными картами существует взаимосвязь - все они являются разными раскрасками одних и тех же нейронов. Каждый пример из обучающей выборки имеет одно и то же расположение на всех картах.

Визуализация деятельности российских банков

Исходные данные.

1800 банков.

Каждый банк описывается 30 финансовыми показателями.

Задача

Построить 2 обобщенных параметра, представляющие данные о банках с минимальной потерей информации.

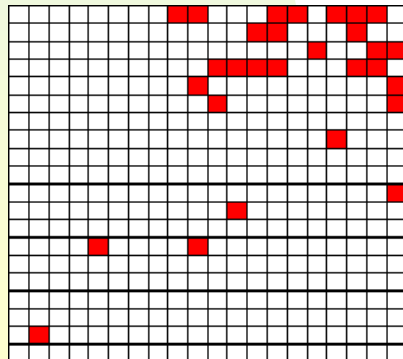
Метод решения: карта Кохонена

Входной слой: 30 нейронов.

Выходной слой: плоскость 20x20 нейронов.

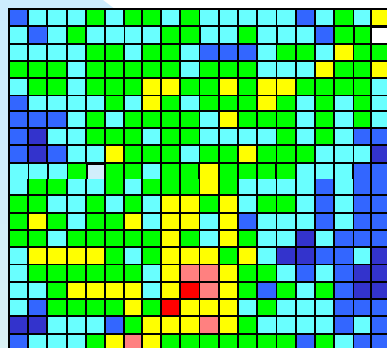
[illegible]

1988. АКБ "ЭЛЕКТРОНИКА"
1571. КБ "ЭРЭП"
1975. АЛТАЙЭНЕРГОБАЛК
2861. КБ "ПИРАНИЯ"
2936. КБ "ИСТАМ"

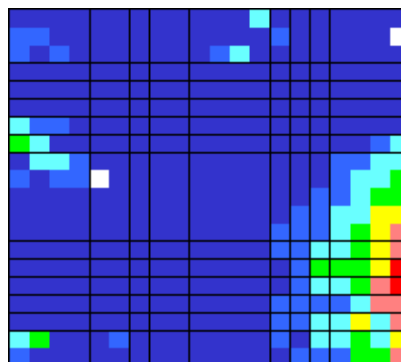


Ячейки, содержащие хотя бы один банк с отозванной лицензией.

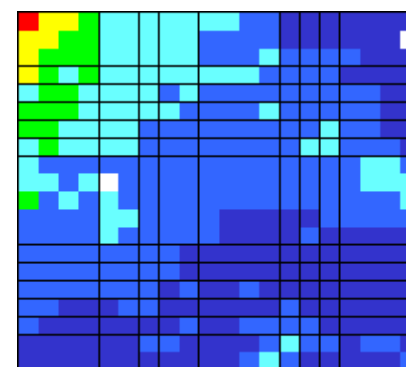
Анализ карт Кохонена



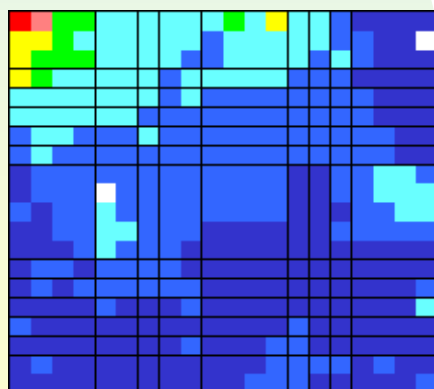
Размер активов



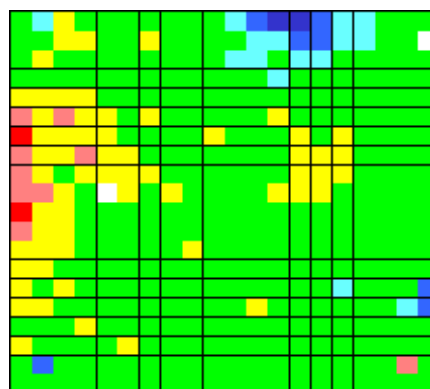
Уставной фонд
(относительный
размер)



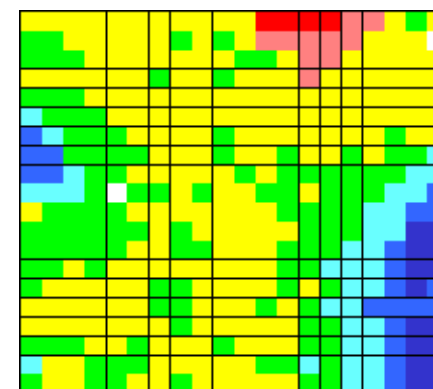
Всего доходов



Всего расходов



Прибыль/убыток



Всего обязательств

- Программное обеспечение, позволяющее работать с картами Кохонена, сейчас представлено множеством инструментов. Это могут быть как инструменты, включающие только реализацию метода самоорганизующихся карт, так и нейропакеты с целым набором структур нейронных сетей, среди которых - и карты Кохонена; также данный метод реализован в некоторых универсальных инструментах анализа данных.
- К инструментарию, включающему реализацию метода карт Кохонена, относятся:
 - ◆ SoMine,
 - ◆ Statistica,
 - ◆ NeuroShell,
 - ◆ NeuroScalp,
 - ◆ Deductor и др.

Достоинства и недостатки

Достоинства:

1. Быстрое обучение;
2. Устойчивость к помехам.

Недостатки:

1. Не обладает свойствами аппроксиматора;
2. Решение задач кластеризации только при известном числе кластеров.