Математические и инструментальные методы машинного обучения

6. Кластеризация

Кластерный анализ

Кластерный анализ (кластер в переводе с лат. означает скопление или гроздь) является совокупностью методов позволяющих исследователю производить т.н. классификацию «без учителя», т.е. разбивать множества объектов на группы, сходные по свойствам, не имея исходных представлений о структуре таких групп.

Результатом работы методов кластерного анализа обычно являются вектор принадлежности объектов к кластерам, таблица объёмов кластеров (количества объектов, содержащихся внутри кластеров) и в некоторых случаях значения критериев качества кластеризации.

Задача кластеризации, т.е. разбиения на группы является субъективной и не всегда имеет «истинное решение», таким образом количество, размер (объём), состав и форма кластеров может меняться для одной и той же задачи при условии применения разных методов, или одних и тех же методов, но с разными параметрами.

Виды задач кластеризации

Методы кластеризации часто используются в прикладных задачах обработки данных, снижения размерности и классификации в маркетинге, социологии, психологии, биологии, в задачах анализа Больших Данных (Big Data), возникающих в КИС (Корпоративных Информационных Системах), установленных на крупных предприятиях, или возникающих в глобальных интернет-проектах (Google, Amazon, Yandex). Можно выделить следующие виды задач:

- Изучение данных
- Облегчение анализа
- Сжатие данных
- Прогнозирование
- Обнаружение аномалий

Постановка задачи кластерного анализа

Исходные данные кластерного анализа представляют собой матрицу N объектов, измеренных по M показателям ($M << N$) или матрицу расстояний (сходств) между N объектами. Каждый объект можно рассматривать как М-мерный вектор значений характеристик.

Матрица расстояний означает набор значений абстрактных расстояний ρ или сходств φ, являющихся мерой совпадения объектных векторов. Задачей кластеризации является разбиение исходной совокупности объектов на кластеры, такие объекты внутри кластера больше похожи друг на друга, чем на объекты из других кластеров.

Таким образом, задачу можно сформулировать следующим образом:

Пусть задано множество наблюдений $X = \{X1, \ldots, XN\}$, Требуется разбить выборку на непересекающиеся подмножества – кластеры $S1, \ldots, Sk$, так чтобы найти экстремум некоторого критерия F, т.е.:

$$S = \{S1, \ldots, Sk\}: F(S) \rightarrow \min (\max) \qquad S$$

Классификация методов кластерного анализа

- Иерархические методы
  - Агломеративные методы
  - Дивизимные методы
- Последовательные методы
  - Методы на основе алгоритма K-средних
  - Плотностные алгоритмы
- Нейгосетевые методы
  - Нейронные сети и карты Кохонена

Меры расстояния и сходства

Одной из принципиальных проблем кластерного анализа является способ определения сходства между объектами. От этого выбора может зависеть форма кластеров, их состав, сходимость алгоритма кластеризации, и возможность интерпретации полученного решения. Считается, что выбор метрики сходства или расстояния надо осуществлять исходя из структуры исходных данных. Метрики сходства и расстояния отличаются тем, что чем более объекты похожи друг на друга, тем сходство выше, а значение расстояния, наоборот, ниже.

$$\rho(X,Y)=\sqrt{\Sigma(X_i-Y_i)^2}$$ Метрика Евклида

$$\rho(X,Y)=\sum_i|X_i-Y_i|$$ Метрика Хемминга

$$\rho(X,Y)=\sqrt[r]{\sum_i|X_i-Y_i|^r}$$ Метрика Минковского

$$\varphi(X,Y)=r_{XY}=\frac{\mathrm{cov}_{XY}}{\sigma_X\sigma_Y}$$ Коэффициент корреляции Пирсона

Бинарные переменные (дихотомические)

- Бинарные переменные обладают только двумя состояниями: 0 и 1;
- Для симметричных бинарных переменных каждое состояние равноценно: например, мужской и женский пол.
- Для несимметричных переменных между состояниями существует разница, например, положительный и отрицательные результаты для исследования наличия заболевания.

Меры сходства для бинарных переменных

- **Для симметричных бинарных переменных:**

$$d(i,j)=\frac{b+c}{a+b+c+d}$$

- **Для несимметричных переменных:**

$$d(i,j)=\frac{b+c}{a+b+c}$$

$$sim_{Jaccard}(i,j)=\frac{a}{a+b+c}$$

		Object j		
		1	0	*sum*
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

Меры сходства для категориальных переменных

Если p – количество признаков, а m – количество совпадений значений признаков:

$$d(i,j)=\frac{p+m}{p}$$

Пример данных для задачи кластеризации. Бинарные переменные

№	пол	возраст	доход	образование
1	0	0	0	0
2	0	1	1	1
3	0	1	1	1
4	1	1	1	1
5	1	0	0	0
6	1	0	0	0

- Пол (0 – муж., 1 – жен.)– симметричная переменная
- Возраст, доход, образование – асимметричные переменные

Метод иерархической агломерации

Метод иерархической агломерации заключается в последовательном объединении N исходных объектов до момента, пока все они не будут объединены в один кластер объёма N. С учётом того, что на каждом шаге подвергаются слиянию только два кластера, процедура содержит N-1 шагов объединения.

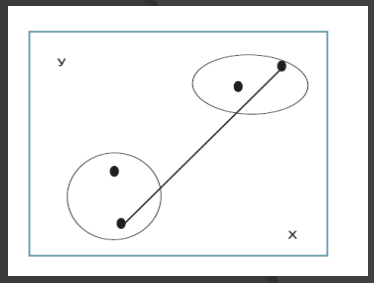
В разных алгоритмах решение о слиянии принимается по-разному, и это определяет форму получаемого кластерного решения. Для вычисления объединяемой пары рассчитывается матрица расстояний между объектами. Расстояние объединения называется расстоянием агломерации и заносится на график, характеризующий процесс работы алгоритма.

Метод иерархической агломерации

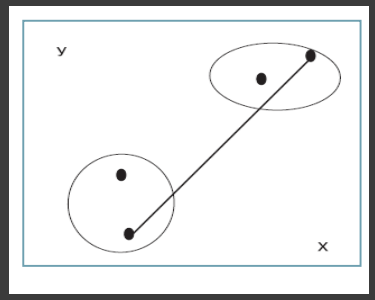
Весь процесс объединения изображают в виде дендрограммы, графика, на котором по оси абсцисс нанесены номера объектов, по оси ординат изображено расстояние объединения. Данный график показывает состав кластерного решения на каждом шаге объединения.

График изменения расстояния агломерации и дендрограмма позволяют принять решение об оптимальной численности кластеров и их составе. Обычно оставлют решение на шаге, после которого расстояние агломерации резко возрастает. Основной недостаток метода в большой вычислительной сложности, которая не подходит для больших выборок и отсутствие наглядности дендрограммы в этом случае.

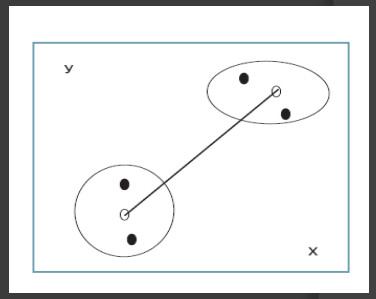
Способы объединения кластеров



Метод «ближайшего соседа» — объединяются кластеры с наименьшим расстоянием между элементами



Метод «дальнего соседа» — объединяются кластеры с наибольшим расстоянием между элементами



Метод центроидов — объединяются кластеры с наименьшим расстоянием между центрами кластеров

Метод Уорда (Варда, Ward) — объединяются кластеры, дающие наименьшую суммарную дисперсию

Метод иерархической агломерации

- **Метод ближайшего соседа (простого связывания)**

Тяготеет к созданию удлиненных, «ленточных» кластеров, вытягивающихся за счет присоединения ближайшей точки. Менее чувствителен к выбросам. Кроме того, метод связан с некоторыми замечательными формальными математическими свойствами, которые, правда, не имеют существенного практического значения.

- **Метод самого дальнего соседа (полного связывания)**

Менее чувствителен к выбросам.

Метод иерархической агломерации

- **Метод центроида**

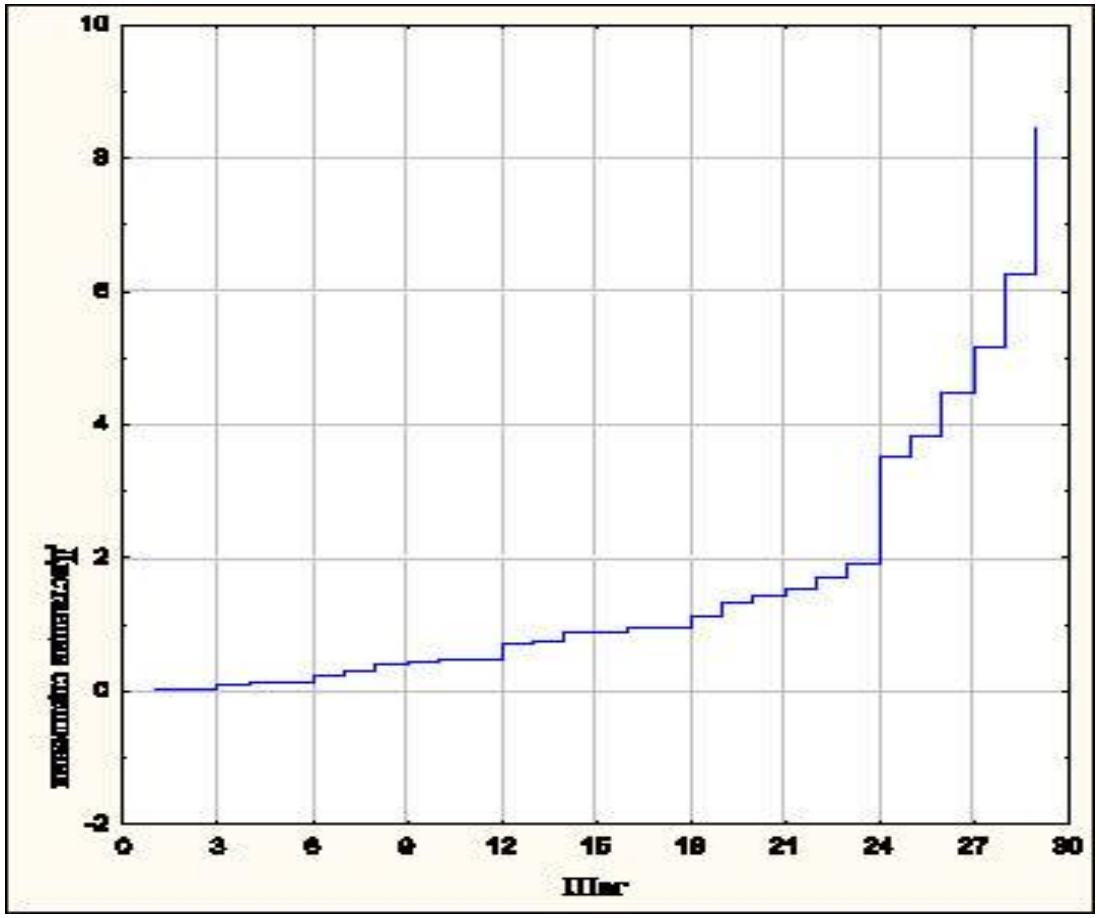
Работает лучше на «засоренных» данных. Менее чувствителен к выбросам.

- **Метод Варда**

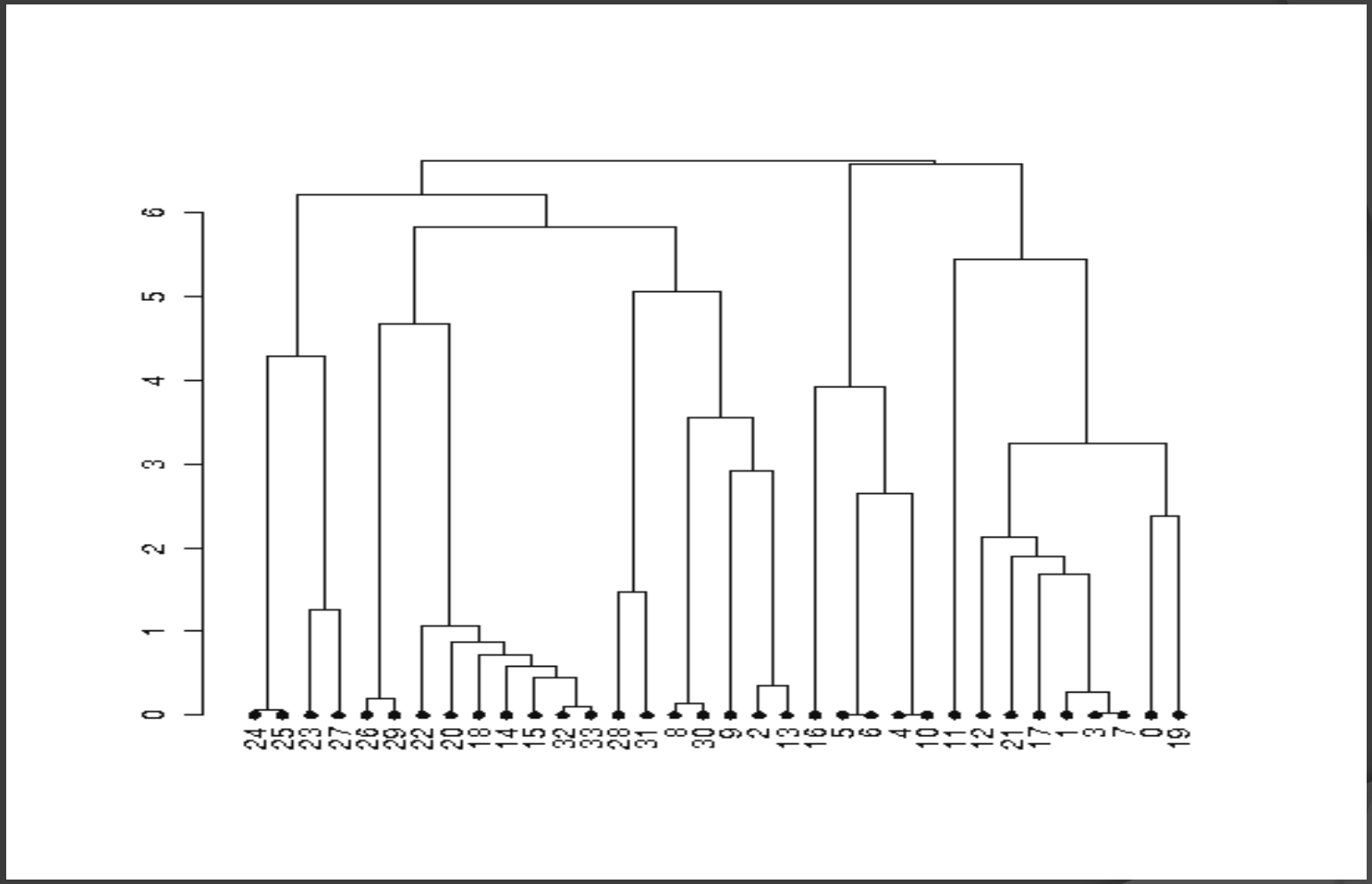
Работает лучше на «засоренных» данных. Чувствителен к выбросам. Возможно образование кластеров со схожими размерами.

Ни один из этих методов не превосходит остальные, каждый имеет свои достоинства и недостатки. Основываясь на исследованиях методом Монте-Карло и эмпирических соображениях, специалисты утверждают, что методы межгруппового среднего связывания, Варда и самого дальнего соседа предпочтительнее остальных.

График агломерации



Дендрограмма



Метод k-средних

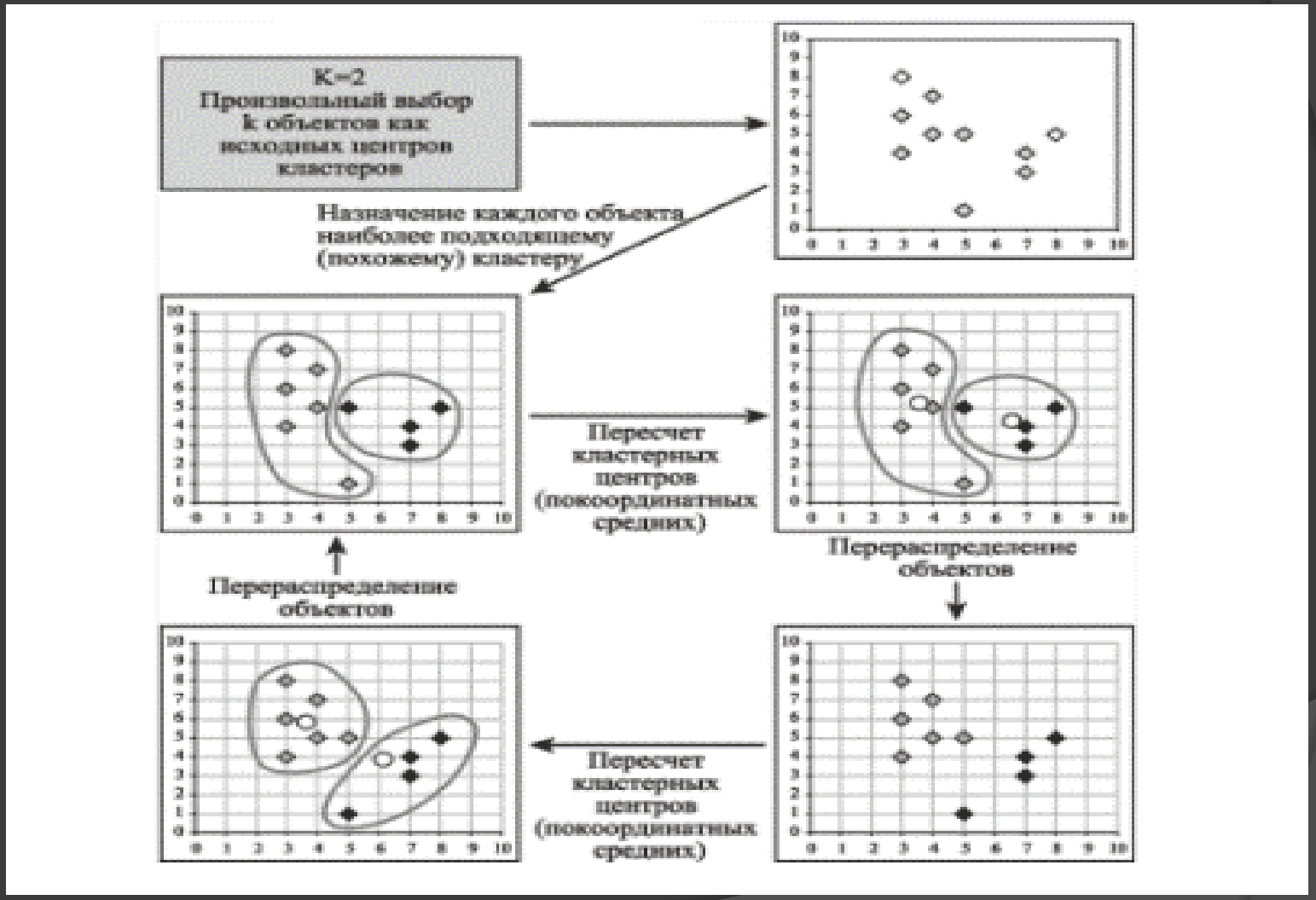
Метод k-средних или метод Мак-Кина заключается в разбиении всего исходного множества объектов вокруг заданных на первом этапе начальных центров k кластеров. Эти центры могут быть получены как с помощью специальных процедур, на основе априорных предположений, так и извлечены из выборки случайным образом.

На каждом шаге итерационного алгоритма из выборки извлекаются объекты и относятся к ближайшему кластеру (обычно используется центроидный метод вычисления расстояния между кластерами). Центроиды пересчитываются, пока не будет выполнено заданное количество итераций, или центроиды не перестанут изменяться.

Метод k-средних

Данный алгоритм имеет множество вариаций и является наиболее распространённым инструментов кластеризации выборов разных размеров, от малых (десятки объектов) до больших (тысячи объектов). Причём, как самомостоятельно, так и в качестве элемента сложных многоэтапных алгоритмов. Основная проблема при использовании метода k-средних лежит в плоскости выбора количества кластеров и их начальных центров.

Метод k-средних



Нейронные сети Кохонена

Сети, называемые картами Кохонена, - это одна из разновидностей нейронных сетей, использующих неконтролируемое обучение. Идея сети Кохонена принадлежит финскому ученому Тойво Кохонену (1982 год). Основной принцип работы сетей - введение в правило обучения нейрона информации относительно его расположения.

Самоорганизующиеся карты могут использоваться для решения таких задач, как моделирование, прогнозирование, поиск закономерностей в больших массивах данных, выявление наборов независимых признаков и сжатие информации. Наиболее распространенное применение сетей Кохонена - решение задачи классификации без учителя, т.е. кластеризации.

Нейронные сети Кохонена

Сеть Кохонена способна распознавать кластеры в данных, а также устанавливать близость классов. Таким образом, пользователь может улучшить свое понимание структуры данных, чтобы затем уточнить нейросетевую модель. Если в данных распознаны классы, то их можно обозначить, после чего сеть сможет решать задачи классификации. Сети Кохонена можно использовать и в тех задачах классификации, где классы уже заданы, - тогда преимущество будет в том, что сеть сможет выявить сходство между различными классами.

Нейронные сети Кохонена

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.

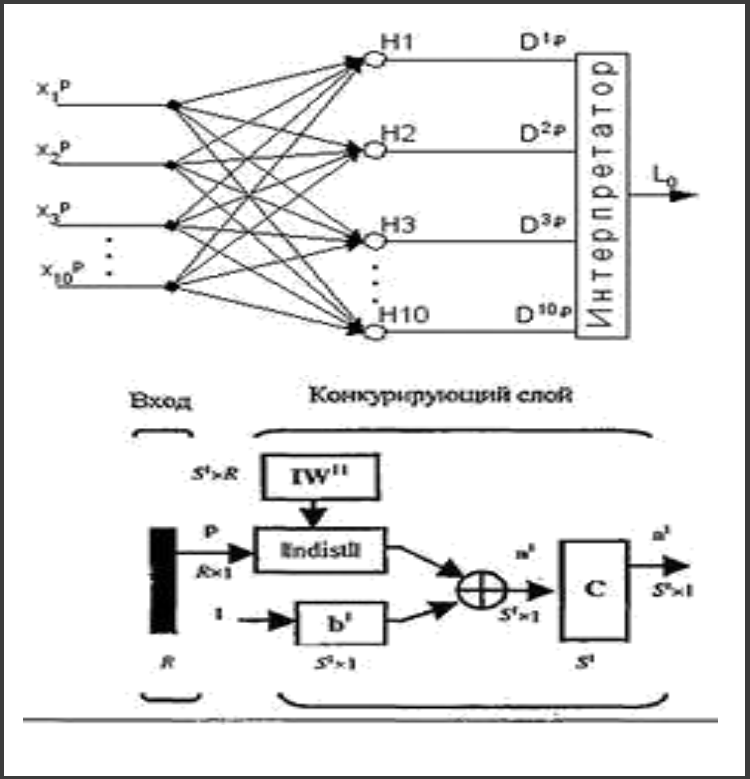
В процессе последовательной подачи на вход сети обучающих примеров определяется наиболее схожий нейрон. Этот нейрон объявляется победителем и является центром при подстройке весов у соседних нейронов. Такое правило обучения предполагает "соревновательное" обучение с учетом расстояния нейронов от "нейрона-победителя".

Нейронные сети Кохонена

Обучение при этом заключается не в минимизации ошибки, а в подстройке весов - внутренних параметров нейронной сети для наибольшего совпадения с входными данными. 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.

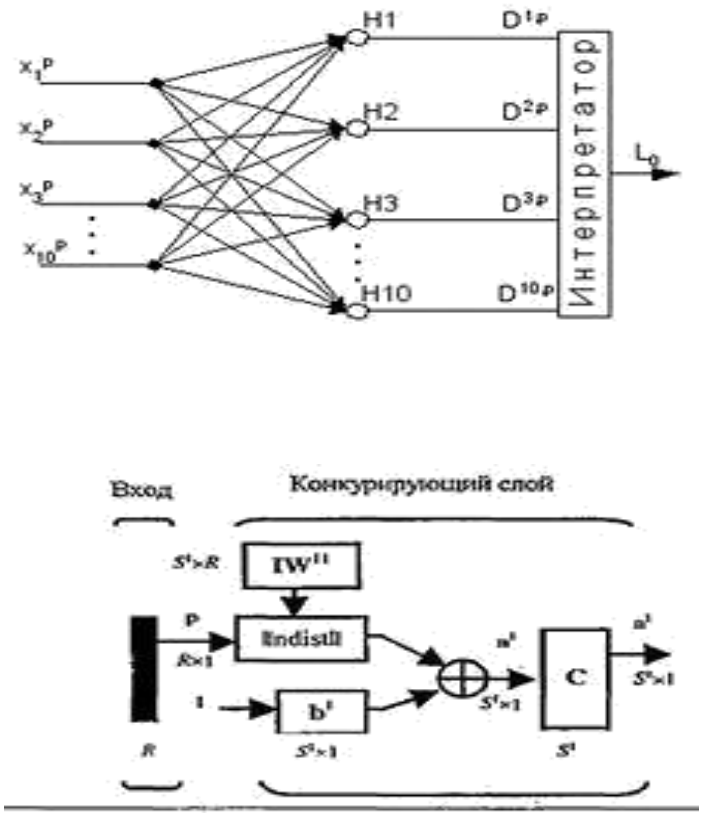
Нейронные сети Кохонена



Самоорганизующаяся карта состоит из компонентов, называемых узлами или нейронами. Их количество задаётся аналитиком. Каждый из узлов описывается двумя векторами. Первый — т. н. вектор веса m, имеющий такую же размерность, что и входные данные. Второй — вектор r, представляющий собой координаты узла на карте. Обычно узлы располагают в вершинах регулярной решётки с квадратными или шестиугольными ячейками.

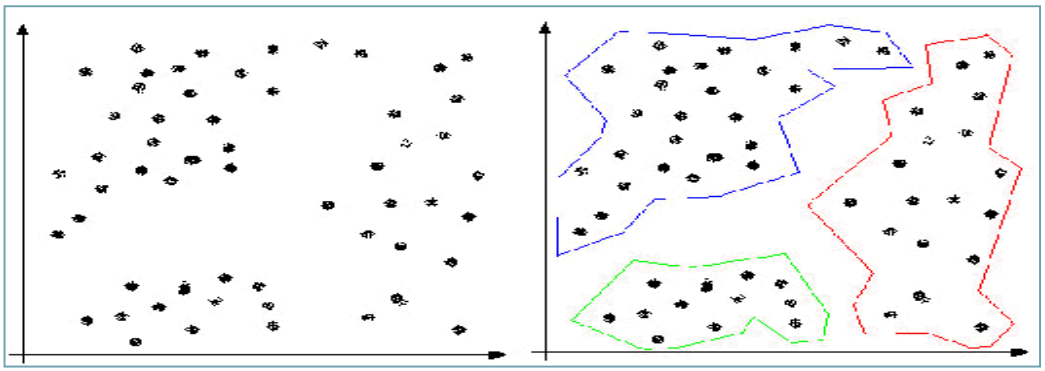
Изначально известна размерность входных данных, по ней некоторым образом строится первоначальный вариант карты. В процессе обучения векторы веса узлов приближаются к входным данным. Для каждого наблюдения (семпла) выбирается наиболее похожий по вектору веса узел, и значение его вектора веса приближается к наблюдению.

Нейронные сети Кохонена

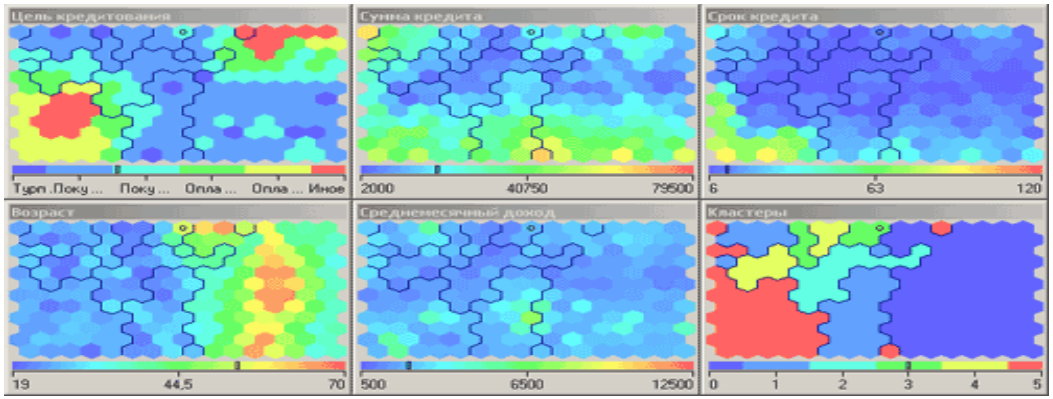


Также к наблюдению приближаются векторы веса нескольких узлов, расположенных рядом, таким образом если в множестве входных данных два наблюдения были схожи, на карте им будут соответствовать близкие узлы. Циклический процесс обучения, перебирающий входные данные, заканчивается по достижении картой допустимой (заранее заданной аналитиком) погрешности, или по совершении заданного количества итераций.

Карты Кохонена

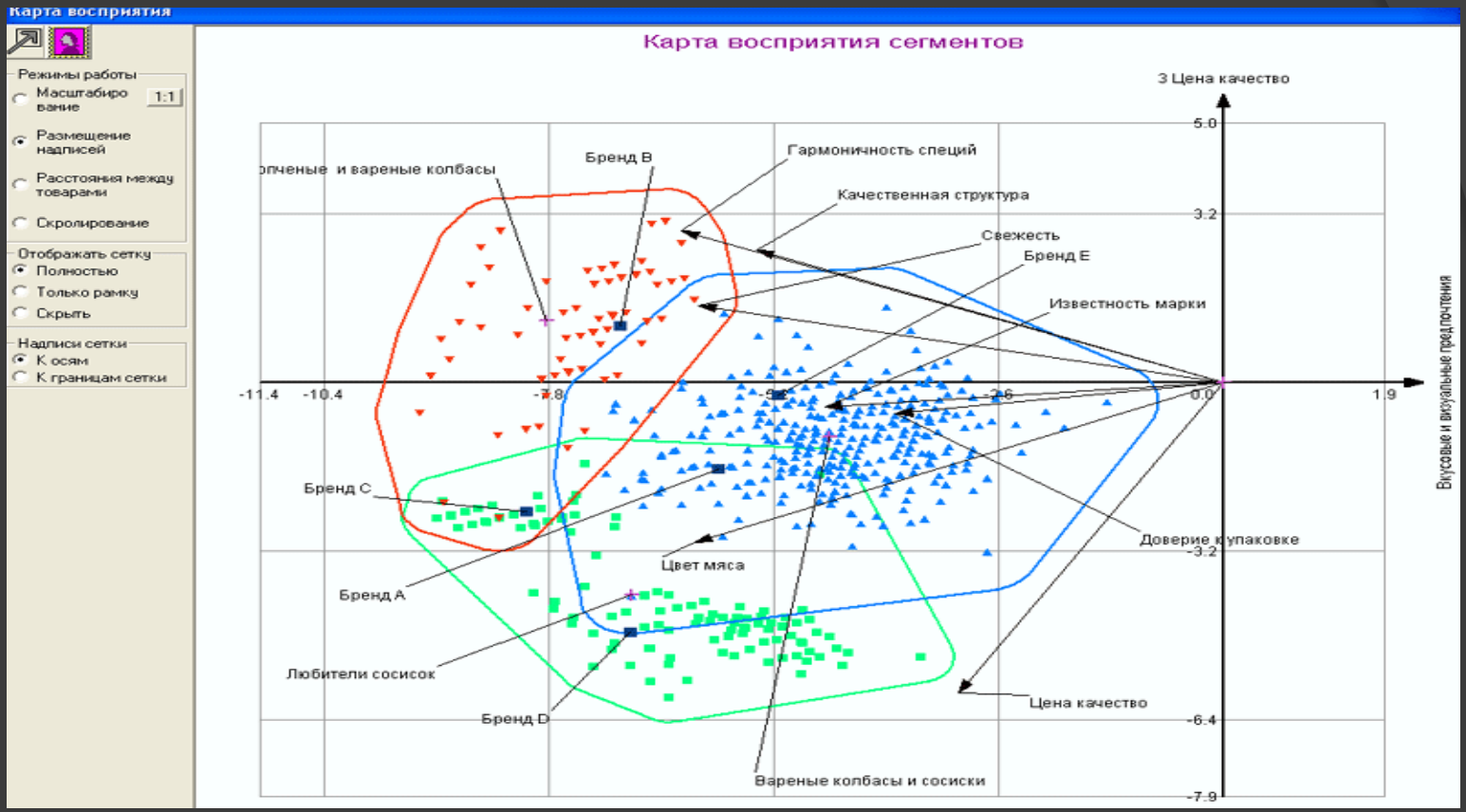


Схематичное представление карты Кохонена



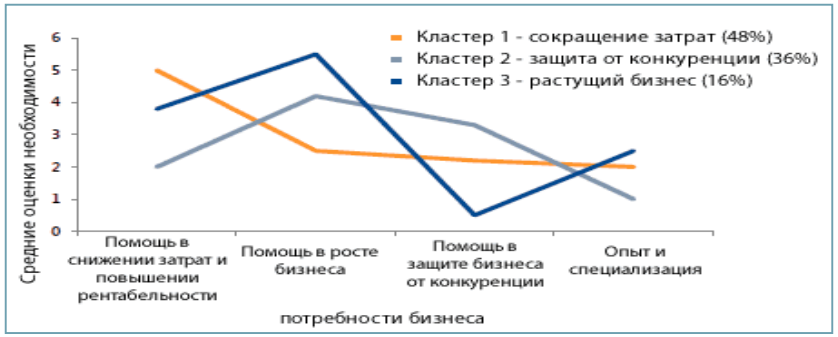
Пример карты Кохонена с раскраской кластеров

Позиционирование брендов



http://ego.uapa.ru/issue/2012/03/04/

Сегментация потребностей бизнеса



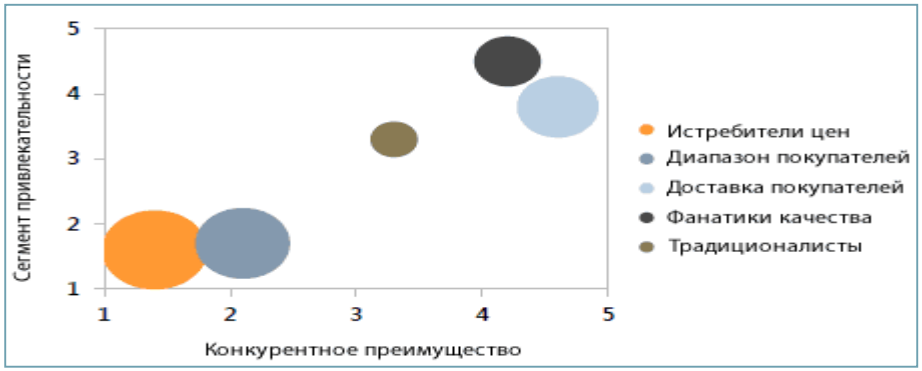
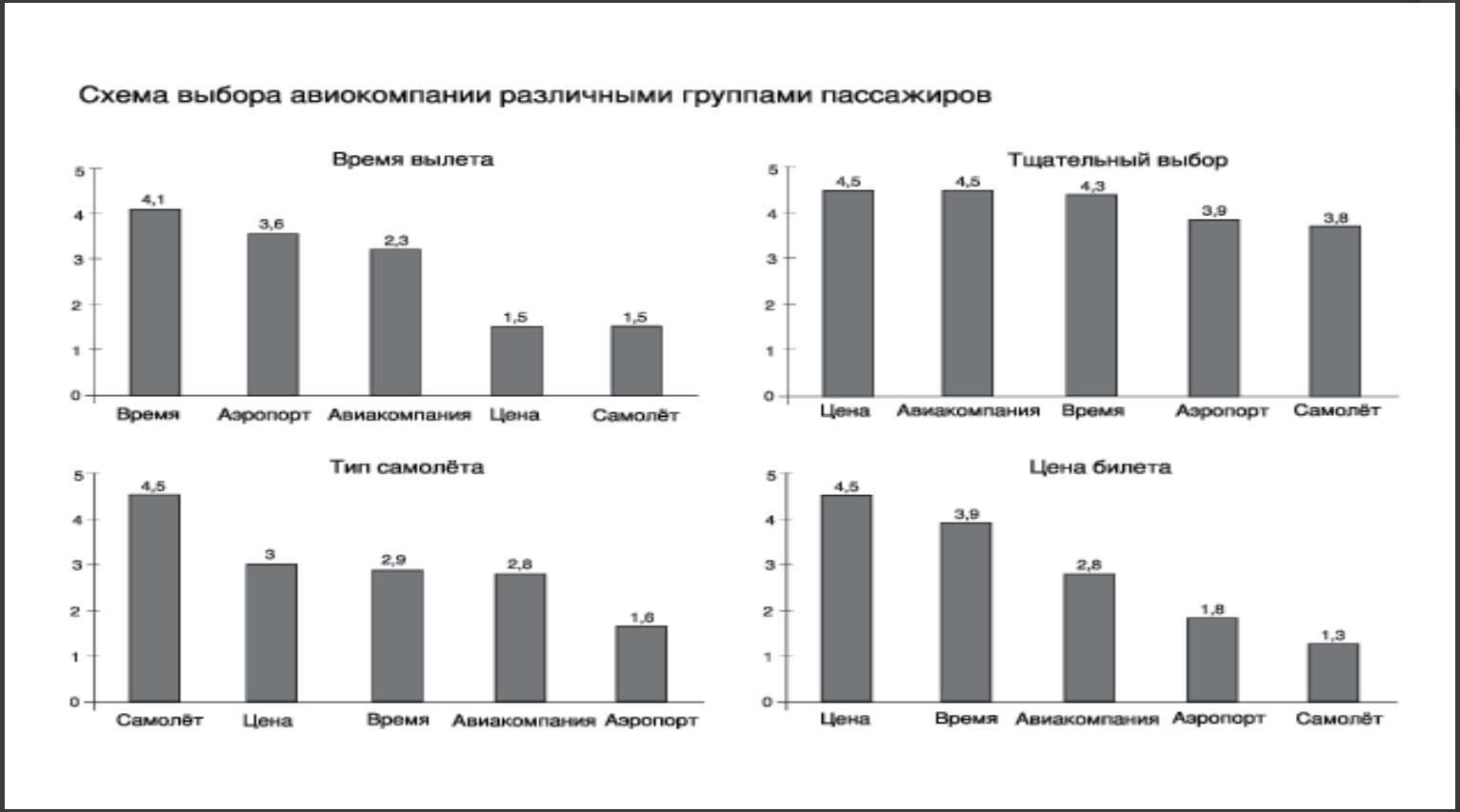


Схема выбора пассажирами авиакомпании-перевозчика



http://www.scanmarket.ru/services/smid18