Методы анализа данных

Деменков П.С.,
Иванисенко В.А.

Понятие образа

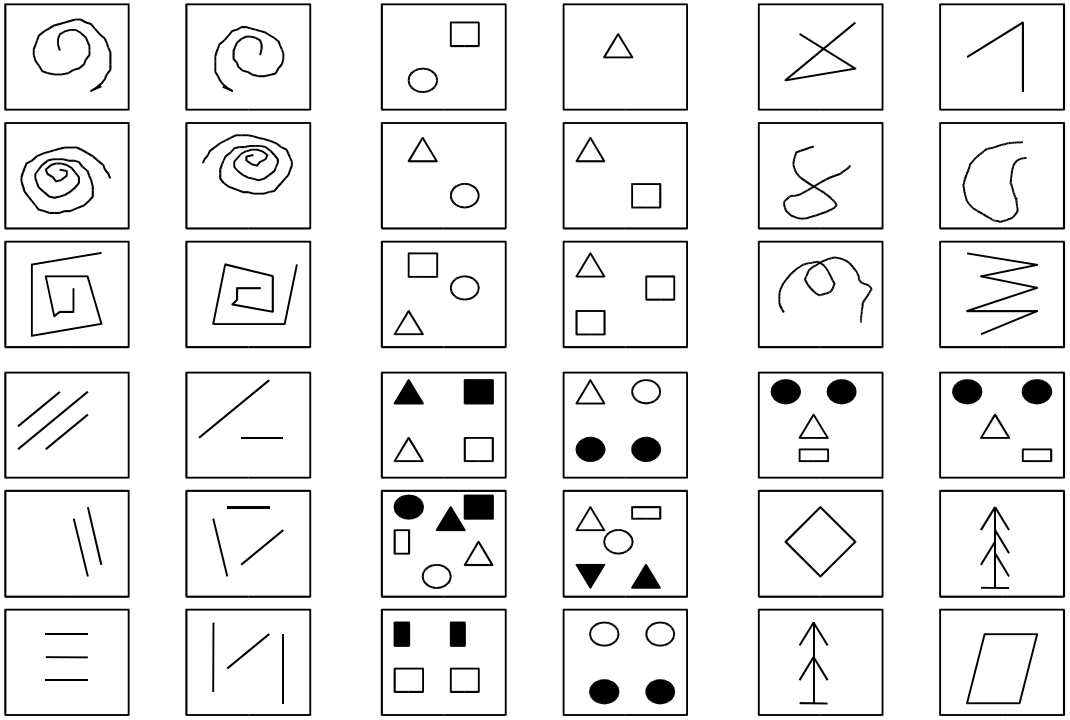
Образ, класс — классификационная группировка в системе, объединяющая (выделяющая) определенную группу объектов по некоторому признаку.

Образы обладают характерным свойством, проявляющимся в том, что ознакомление с конечным числом явлений из одного и того же множества дает возможность узнавать сколь угодно большое число его представителей.

Примерами образов могут быть: река, море, жидкость, и т. д.

Разные люди, обучающиеся на различном материале наблюдений, большей частью одинаково и независимо друг от друга классифицируют одни и те же объекты. Именно эта объективность образов позволяет людям всего мира понимать друг друга.

Проблема обучения распознаванию образов (ОРО)



Требуется отобрать признаки, при помощи которых можно отличить левую триаду картинок от правой. Решение данных задач требует моделирования логического мышления в полном объеме.

Проблема обучения распознаванию образов (ОРО)

В целом проблема распознавания образов состоит из двух частей: обучения и распознавания. Обучение осуществляется путем показа отдельных объектов с указанием их принадлежности тому или другому образу. В результате обучения распознающая система должна приобрести способность реагировать одинаковыми реакциями на все объекты одного образа и различными — на все объекты различных образов. Очень важно, что процесс обучения должен завершиться только путем показов конечного числа объектов без каких-либо других подсказок.

Важно, что в процессе обучения указываются только сами объекты и их принадлежность образу. За обучением следует процесс распознавания новых объектов, который характеризует действия уже обученной системы. Автоматизация этих процедур и составляет проблему обучения распознаванию образов.

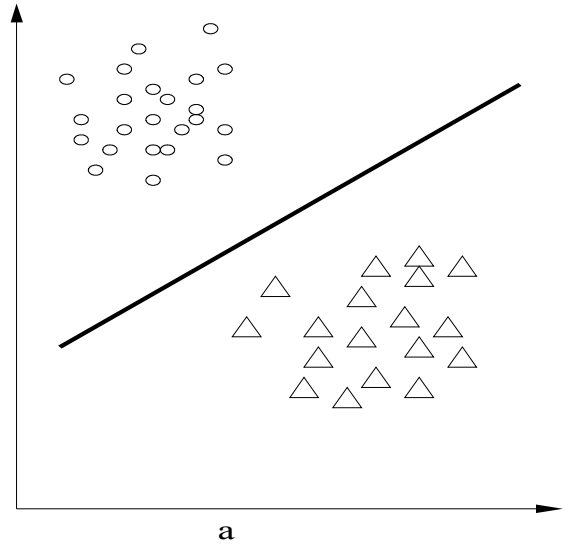
Проблема обучения распознаванию образов (ОРО)

Прежде чем начать анализ какого-либо объекта, нужно получить о нем определенную, каким-либо способом упорядоченную информацию. Такая информация представляет собой характеристику объектов, их отображение на множестве воспринимающих органов распознающей системы.

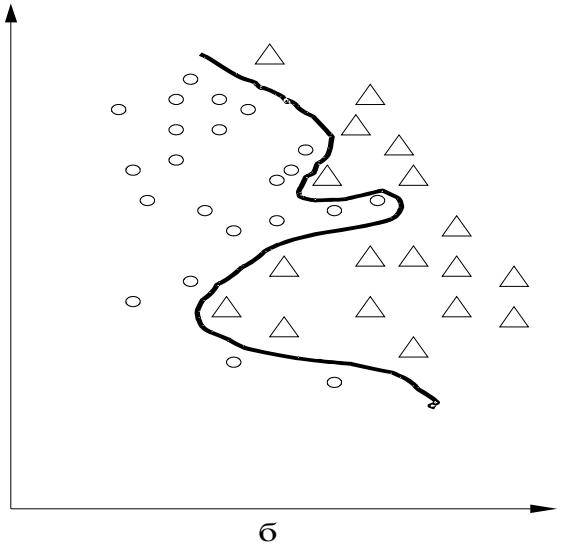
Но каждый объект наблюдения может воздействовать по-разному, в зависимости от условий восприятия. Кроме того, объекты одного и того же образа могут достаточно сильно отличаться друг от друга и, естественно, по-разному воздействовать на воспринимающие органы.

Каждое отображение какого-либо объекта на воспринимающие органы распознающей системы, независимо от его положения относительно этих органов, принято называть изображением объекта, а множества таких изображений, объединенные какими-либо общими свойствами, представляют собой образы

Геометрический подход



а



б

Цель обучения состоит в построении таких функций от векторов-изображений, которые была бы, например, положительны на всех точках одного и отрицательны на всех точках другого образа. В связи с тем, что области не имеют общих точек, всегда существует целое множество таких разделяющих функций, а в результате обучения должна быть построена одна из них.

Структурный (лингвистический) подход

- **Выделяется набор исходных понятий — типичных фрагментов, встречающихся на изображениях, и характеристик взаимного расположения фрагментов — "слева", "снизу", "внутри" и т. д. Эти исходные понятия образуют словарь, позволяющий строить различные логические высказывания, иногда называемые предположениями. Задача — отобрать наиболее существенные для данного конкретного случая высказывания.**
- **Просматривая конечное и по возможности небольшое число объектов из каждого образа, нужно построить описание этих образов. Построенные описания должны быть столь полными, чтобы решить вопрос о том, к какому образу принадлежит данный объект.**

Гипотеза компактности

Если предположить, что в процессе обучения пространство признаков формируется исходя из задуманной классификации, то тогда можно надеяться, что задание пространство признаков само по себе задает свойство, под действием которого образы в этом пространстве легко разделяются

Гипотеза компактности — образам соответствуют компактные множества в пространстве признаков.

Под компактным множеством пока будем понимать некие "сгустки" точек в пространстве изображений, предполагая, что между этими сгустками существуют разделяющие их разряжения.

Процесс самообучения

Процессом самообучения некоторой системы называется такой процесс, в результате которого эта система без подсказки учителя приобретает способность к выработке одинаковых реакций на изображения объектов одного и того же образа и различных реакций на изображения различных образов.

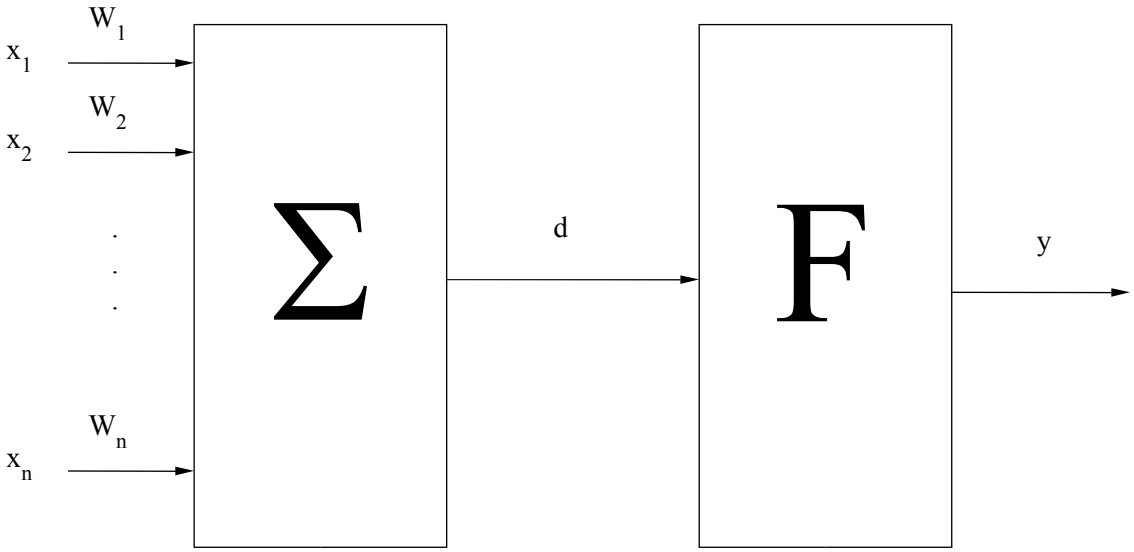
Роль учителя при этом состоит лишь в подсказке системе некоторого объективного свойства, одинакового для всех образов и определяющего способность к разделению множества объектов на образы.

Обучение и адаптация

Обучением называют процесс выработки в некоторой системе той или иной реакции на группы внешних идентичных сигналов путем многократного воздействия на систему внешней корректировки. Такую внешнюю корректировку в обучении принято называть "поощрениями" и "наказаниями". Механизм генерации этой корректировки практически полностью определяет алгоритм обучения.

Адаптация — это процесс изменения параметров и структуры системы, а возможно, и управляющих воздействий на основе текущей информации с целью достижения определенного состояния системы при начальной неопределенности и изменяющихся условиях работы.

Математическая модель нейрона

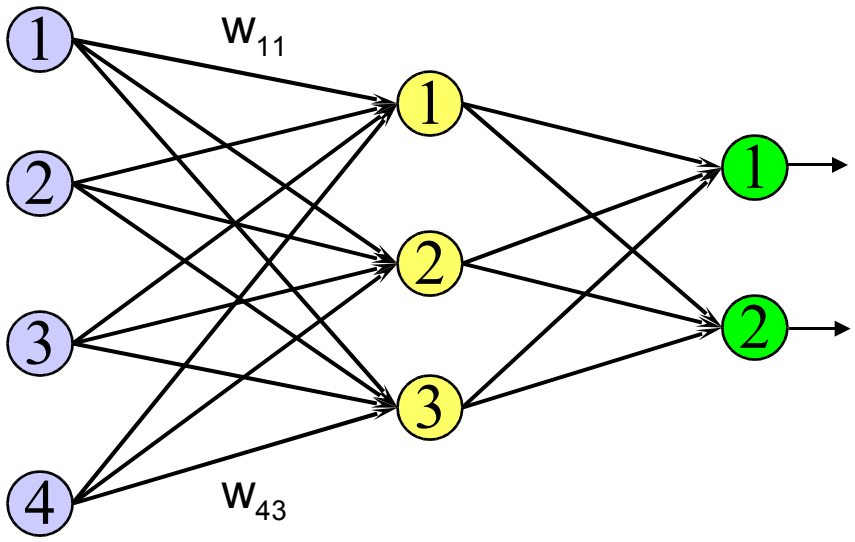


Суммирующий элемент
$d=\Sigma W_i * x_i$

Активационный элемент
$y=F(d)$

- **Будучи соединенными определенным образом, нейроны образуют нейронную сеть. Работа сети разделяется на обучение и адаптацию. Под обучением понимается процесс адаптации сети к предъявляемым эталонным образцам путем модификации (в соответствии с тем или иным алгоритмом) весовых коэффициентов связей между нейронами.**
- **Этот процесс является результатом алгоритма функционирования сети, а не предварительно заложенных в нее знаний человека**

Структура нейронной сети (BPN)

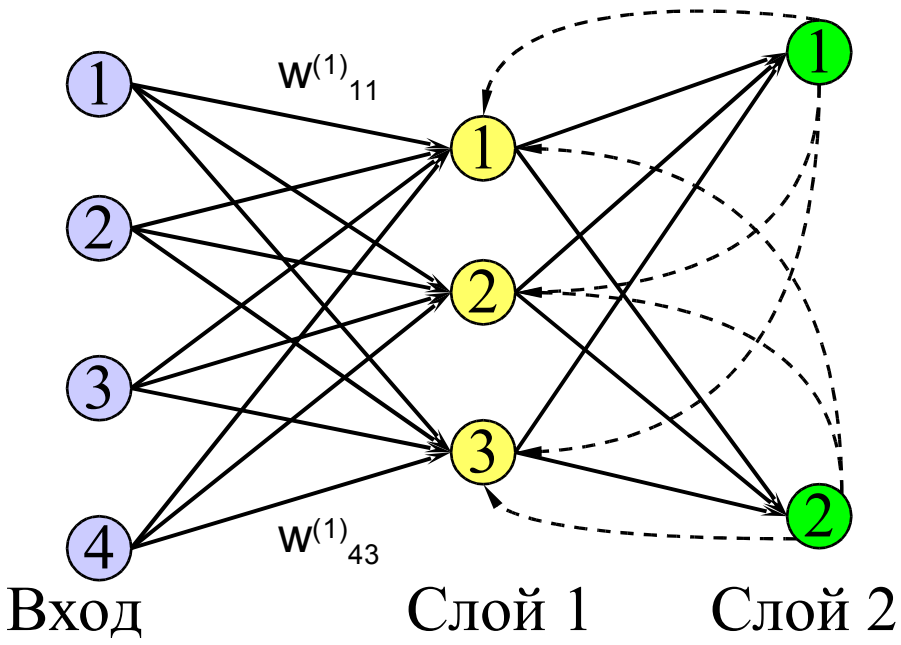


Одной из наиболее известных структур нейронных сетей является многослойная структура, в которой каждый нейрон произвольного слоя связан со всеми аксонами нейронов предыдущего слоя

Способы изменения весов в сети

- **Разработка наборов выходных сигналов, соответствующих входным, для каждого слоя НС**
- **Динамическая подстройка весовых коэффициентов синапсов, в ходе которой выбираются, как правило, наиболее слабые связи и изменяются на малую величину в ту или иную сторону, а сохраняются только те изменения, которые повлекли уменьшение ошибки на выходе всей сети.**
- **Распространение сигналов ошибки от выходов НС к ее входам, в направлении, обратном прямому распространению сигналов в обычном режиме работы.**

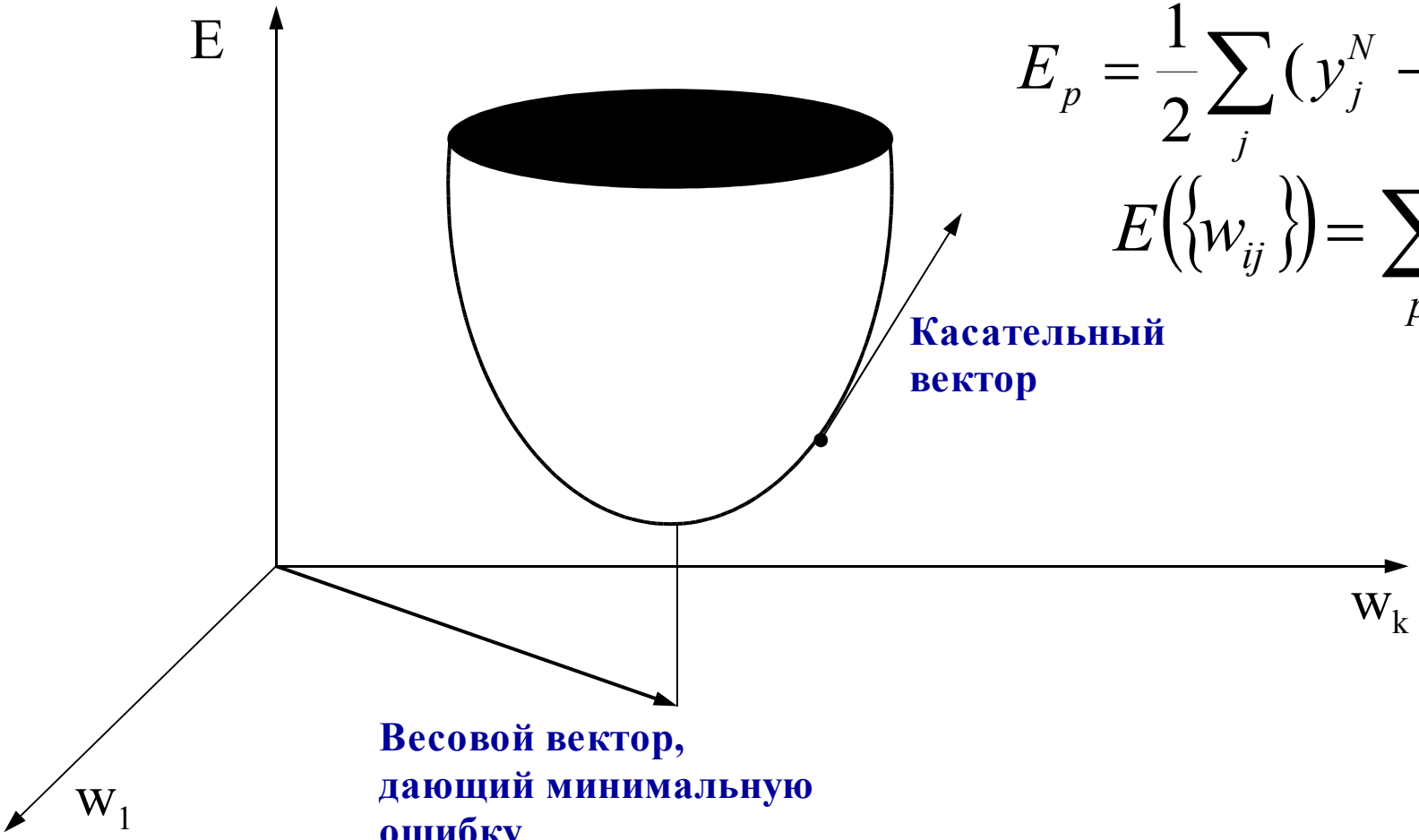
Общий алгоритм работы BPN



- **Прямой проход**
  По заданным значениям входных нейронов сети рассчитываются значения нейронов на выходе

- **Обратный проход**
  Вычисляется ошибка и определяется какой вклад в эту ошибку внес каждый из нейронов сети

Поверхность ошибки



$$E_p = \frac{1}{2}\sum_j (y_j^N - d_j)^2$$

$$E(\{w_{ij}\}) = \sum_p E_p$$

Обучение градиентным спуском

$$\Delta w_{ij}^{(n)} = -\eta \cdot \frac{\partial E}{\partial w_{ij}}$$

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_j} \cdot \frac{dy_j}{ds_j} \cdot \frac{\partial s_j}{\partial w_{ij}}$$

$$\frac{\partial E}{\partial y_j} = \sum_k \frac{\partial E}{\partial y_k} \cdot \frac{dy_k}{ds_k} \cdot \frac{\partial s_k}{\partial y_j} = \sum_k \frac{\partial E}{\partial y_k} \cdot \frac{dy_k}{ds_k} \cdot w_{jk}^{(n+1)}$$

$$\delta_j^{(n)} = \left[\sum_k \delta_k^{(n+1)} \cdot w_{jk}^{(n+1)} \right] \cdot \frac{dy_j}{ds_j}$$

$$\delta_l^{(N)} = (y_l^{(N)} - d_l) \cdot \frac{dy_l}{ds_l}$$

$$f(x) = \frac{1}{1 + e^{-\alpha \cdot x}}$$

$$\Delta w_{ij}^{(n)} = -\eta \cdot \delta_j^{(n)} \cdot y_i^{(n-1)}$$

Оценка количества нейронов в сети

- Для НС с двумя слоями, то есть выходным и одним скрытым слоем, детерминистская емкость (количество распознаваемых образов) сети C_d оценивается так:

$$N_w/N_y<C_d<N_w/N_y \cdot \log(N_w/N_y) ,$$

где N_w – число подстраиваемых весов, N_y – число нейронов в выходном слое.

4. число входов N_x и нейронов в скрытом слое N_h должно удовлетворять неравенству $N_x+N_h>N_y$.
5. $N_w/N_y>1000$.

Мощность выходного слоя сети

- **1 нейрон соответствует 2 классам (0, 1)**
- **2 нейрона соответствуют 4 классам и т.д.**

Однако результаты работы сети, организованной таким образом, можно сказать – "под завязку", – не очень надежны.

Для повышения достоверности классификации желательно ввести избыточность путем выделения каждому классу одного нейрона в выходном слое или, что еще лучше, нескольких, каждый из которых обучается определять принадлежность образа к классу со своей степенью достоверности, например: высокой, средней и низкой. Такие НС позволяют проводить классификацию входных образов, объединенных в нечеткие (размытые или пересекающиеся) множества. Это свойство приближает подобные НС к условиям реальной жизни.

«Узкие места» нейронной сети

- **В процессе обучения может возникнуть ситуация, когда большие положительные или отрицательные значения весовых коэффициентов сместят рабочую точку на сигмоидах многих нейронов в область насыщения.**
- **Применение метода градиентного спуска не гарантирует, что будет найден глобальный, а не локальный минимум целевой функции.**
- **Выбор величины скорости обучения.**

Нейронные сети: обучение без учителя

Главная черта, делающая обучение без учителя привлекательным, – это его "самостоятельность". Процесс обучения, как и в случае обучения с учителем, заключается в подстраивании весов синапсов. Некоторые алгоритмы, правда, изменяют и структуру сети, то есть количество нейронов и их взаимосвязи, но такие преобразования правильнее назвать более широким термином – самоорганизацией.

Сеть Хебба

В 1949г. канадский психолог Д.Хебб опубликовал книгу “Организация поведения” (D.Hebb “Organizational Behaviour”), в которой он постулировал правдоподобный механизм обучения на клеточном уровне в мозге.

Основная идея Хебба состояла в том, что когда входной сигнал нейрона, поступающий через синаптические связи, вызывает срабатывание нейрона, то эффективность такого входа в терминах его способности содействовать срабатыванию нейрона в будущем должна увеличиваться.

Хебб предположил, что изменение эффективности должно происходить именно в синапсе, который подает этот сигнал на вход нейрона назначения. Позднейшие исследования подтвердили эту догадку Хебба. Хотя в последнее время были открыты другие механизмы биологического обучения на клеточном уровне.

Сигнальный метод обучения Хебба

$$w_{ij}(t) = w_{ij}(t-1) + \alpha \cdot y_i^{(n-1)} \cdot y_j^{(n)}$$

где $y_i^{(n-1)}$ – выходное значение нейрона i слоя (n-1), $y_j^{(n)}$ – выходное значение нейрона j слоя n; $w_{ij}(t)$ и $w_{ij}(t-1)$ – весовой коэффициент синапса, соединяющего эти нейроны, на итерациях t и t 1 соответственно; α – коэффициент скорости обучения.

При обучении по данному методу усиливаются связи между возбужденными нейронами.

Существует также и дифференциальный метод обучения Хебба.

$$w_{ij}(t) = w_{ij}(t-1) + \alpha \cdot \left[y_i^{(n-1)}(t) - y_i^{(n-1)}(t-1)\right] \cdot \left[y_j^{(n)}(t) - y_j^{(n)}(t-1)\right]$$

Здесь $y_i^{(n-1)}(t)$ и $y_i^{(n-1)}(t-1)$ – выходное значение нейрона i слоя n-1 соответственно на итерациях t и t-1; $y_j^{(n)}(t)$ и $y_j^{(n)}(t-1)$ – то же самое для нейрона j слоя n.

Полный алгоритм обучения

1. **Всем весовым коэффициентам присваиваются небольшие случайные значения.**
2. **На входы сети подается входной образ, и сигналы возбуждения распространяются по всем слоям согласно принципам классических прямопоточных (feedforward) сетей**
3. **На основании полученных выходных значений нейронов по формуле (1) или (2) производится изменение весовых коэффициентов.**
4. **Цикл с шага 2, пока выходные значения сети не застабилизируются с заданной точностью.**

Анализ алгоритма

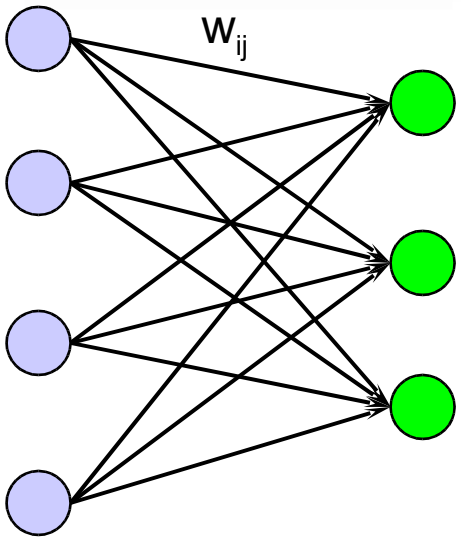
На втором шаге цикла попеременно предъявляются все образы из входного набора.

Следует отметить, что вид откликов на каждый класс входных образов не известен заранее и будет представлять собой произвольное сочетание состояний нейронов выходного слоя, обусловленное случайным распределением весов на стадии инициализации. Вместе с тем, сеть способна обобщать схожие образы, относя их к одному классу. Тестирование обученной сети позволяет определить топологию классов в выходном слое. Для приведения откликов обученной сети к удобному представлению можно дополнить сеть одним слоем, который, например, по алгоритму обучения однослойного персептрона необходимо заставить отображать выходные реакции сети в требуемые образы.

$$w_{ij}(t) = w_{ij}(t-1) + \alpha \cdot \left[y_i^{(n-1)} - w_{ij}(t-1) \right]$$

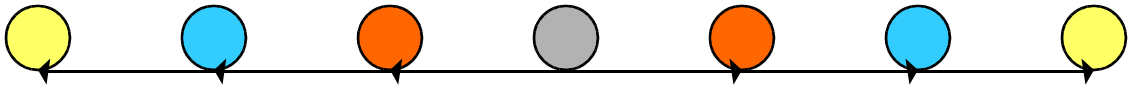
обучение сводится к минимизации разницы между входными сигналами нейрона, поступающими с выходов нейронов предыдущего слоя $\mathbf{y_i^{(n-1)}}$, и весовыми коэффициентами его синапсов.

Алгоритм работы сети Кохонена



$$d_j = \sum_i (w_{ij} - x_i)^2$$

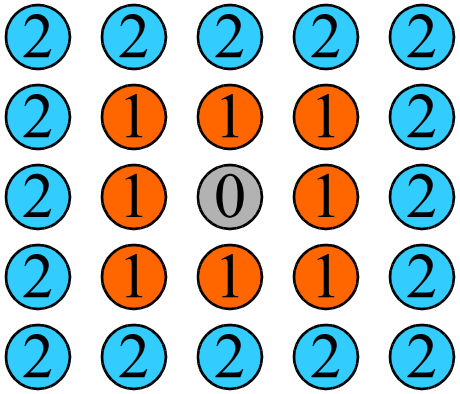
Варианты определения радиуса



Радиус 1

Радиус 2

Радиус 3



Замечания по алгоритму

Инициализация весовых коэффициентов случайными значениями может привести к тому, что различные классы, которым соответствуют плотно распределенные входные образы, сольются или, наоборот, раздробятся на дополнительные подклассы в случае близких образов одного и того же класса. Для обхода такой ситуации используется метод выпуклой комбинации[3]. Суть его сводится к тому, что входные нормализованные образы подвергаются преобразованию:

$$x_i = \alpha(t) \cdot x_i + (1 - \alpha(t)) \cdot \frac{1}{\sqrt{n}}$$

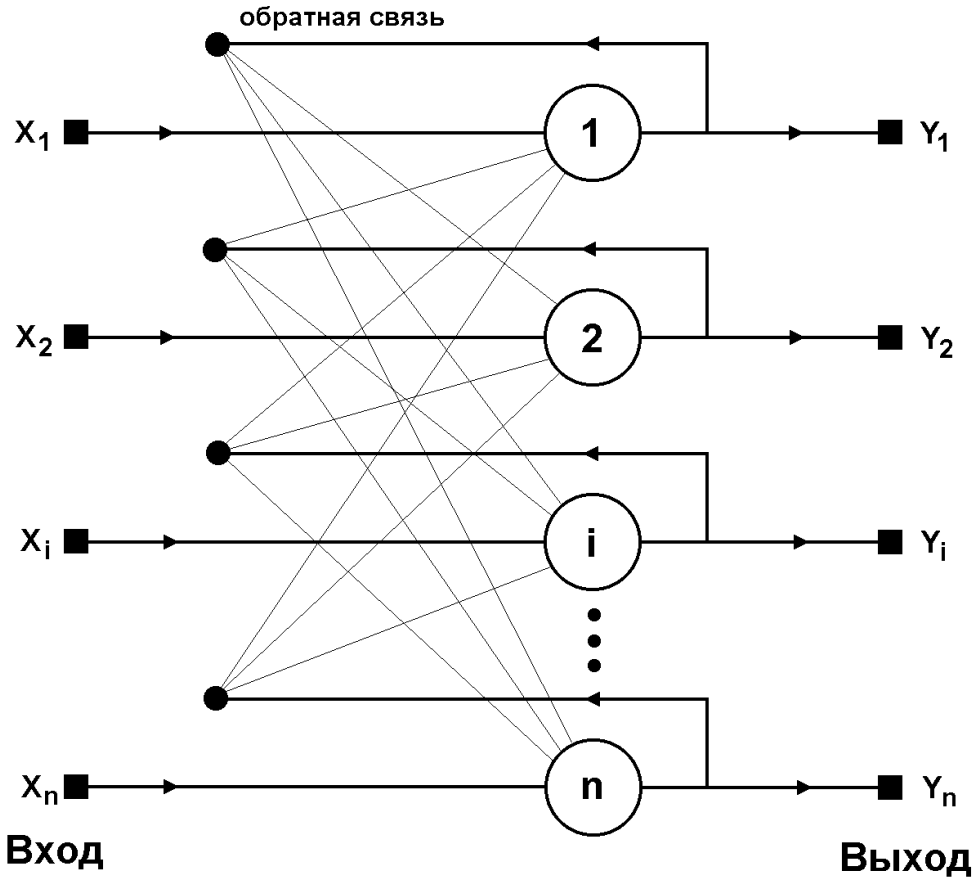
Весовые коэффициенты устанавливаются на шаге инициализации равными величине

$$w_o = \frac{1}{\sqrt{n}}$$

Сети ассоциативной памяти

Среди различных конфигураций искусственных нейронных сетей (НС) встречаются такие, при классификации которых по принципу обучения, строго говоря, не подходят ни обучение с учителем, ни обучение без учителя. В таких сетях весовые коэффициенты синапсов рассчитываются только однажды перед началом функционирования сети на основе информации об обрабатываемых данных, и все обучение сети сводится именно к этому расчету. С одной стороны, предъявление априорной информации можно расценивать, как помощь учителя, но с другой – сеть фактически просто запоминает образцы до того, как на ее вход поступают реальные данные, и не может изменять свое поведение, поэтому говорить о звене обратной связи с "миром" (учителем) не приходится. Из сетей с подобной логикой работы наиболее известны сеть Хопфилда и сеть Хэмминга, которые обычно используются для организации ассоциативной памяти.

Структурная схема сети Хопфилда



Задача, решаемая сетью Хопфилда

Известен некоторый набор двоичных сигналов (изображений, звуковых оцифровок, прочих данных, описывающих некие объекты или характеристики процессов), которые считаются образцовыми. Сеть должна уметь из произвольного неидеального сигнала, поданного на ее вход, выделить ("вспомнить" по частичной информации) соответствующий образец (если такой есть) или "дать заключение" о том, что входные данные не соответствуют ни одному из образцов.

Принцип работы

В общем случае входной сигнал может быть описан вектором X = { x_i: i=0...n-1} ($x_i \in \{-1, 1\}$)

X^k - вектор, описывающий k-ый образец

Когда сеть распознает (или "вспомнит") какой-либо образец на основе предъявленных ей данных, ее выходы будут содержать именно его, то есть Y = X^k, где Y – вектор выходных значений сети. В противном случае, выходной вектор не совпадет ни с одним образцовым.

На стадии инициализации сети весовые коэффициенты синапсов устанавливаются следующим образом:

$$w_{ij} = \begin{cases} \sum_{k=0}^{m-1} x_i^k x_j^k, i \neq j \\ 0, \qquad i = j \end{cases}$$

Алгоритм работы сети

1. На входы сети подается неизвестный сигнал. Фактически его ввод осуществляется непосредственной установкой значений аксонов:

$y_i(0) = xi$, $i = 0...n-1$,

поэтому обозначение на схеме сети входных синапсов в явном виде носит чисто условный характер. Ноль в скобке справа от yi означает нулевую итерацию в цикле работы сети.

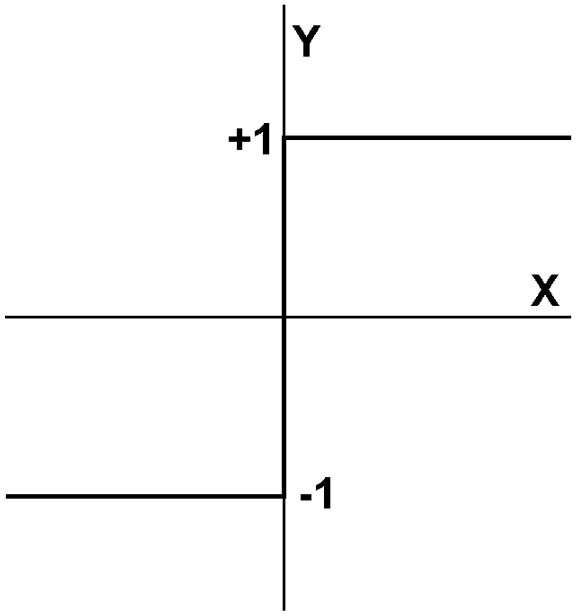
2. Рассчитывается новое состояние нейронов

$$s_j(p+1) = \sum_{i=0}^{n-1} w_{ij} y_i(p)$$

и новые значения аксонов

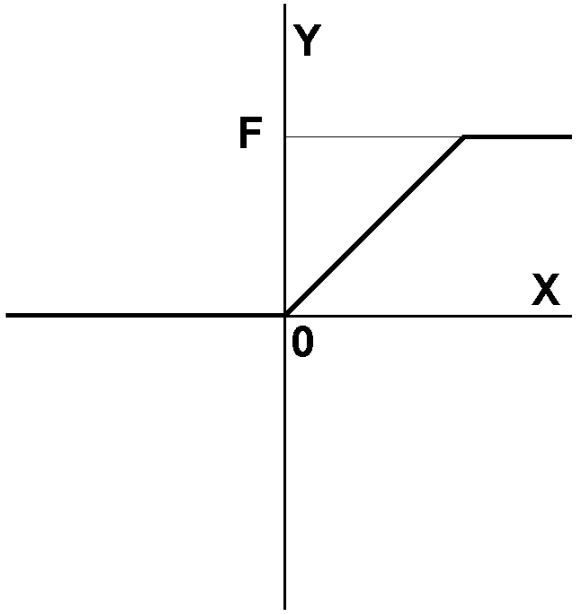
$$y_j(p+1) = f\left[s_j(p+1)\right]$$

Активационные функции.



а)

Для сети Хопфилда



б)

Для сети Хэмминга

Алгоритм работы сети

3. Проверка, изменились ли выходные значения аксонов за последнюю итерацию. Если да – переход к пункту 2, иначе (если выходы застабилизировались) – конец. При этом выходной вектор представляет собой образец, наилучшим образом сочетающийся с входными данными.

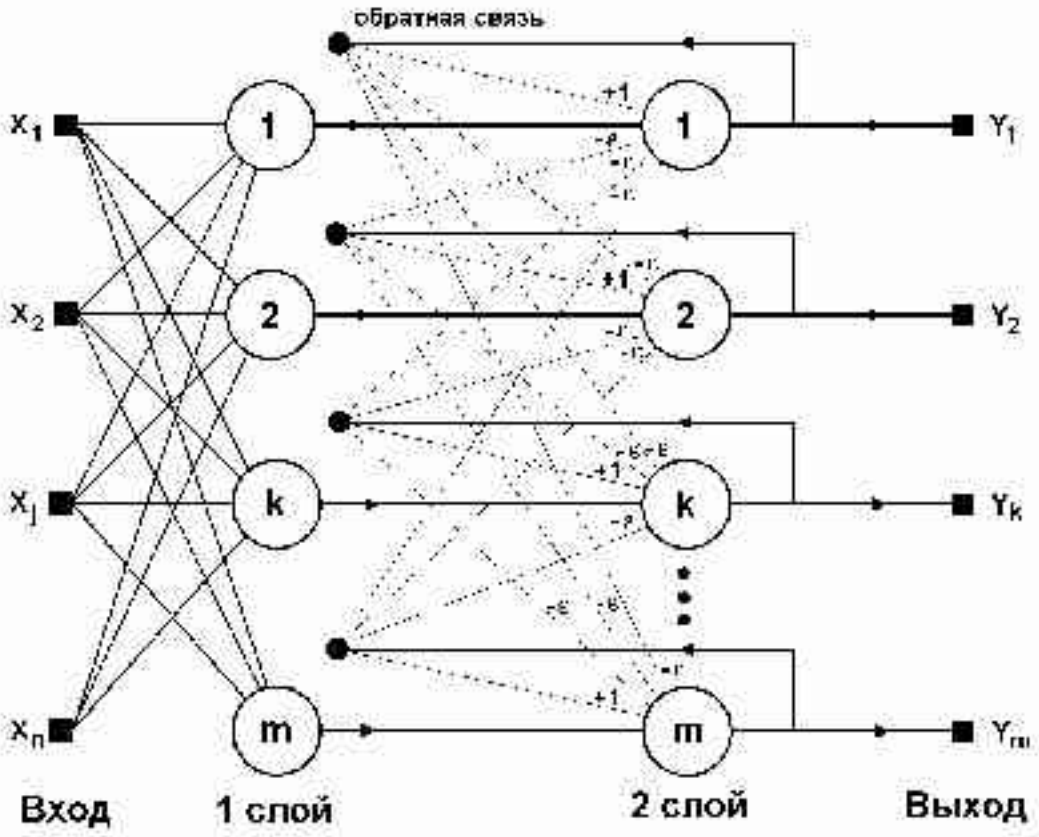
Ограничения сети

Сеть не может провести распознавание и выдает на выходе несуществующий образ. Это связано с проблемой ограниченности возможностей сети. Для сети Хопфилда число запоминаемых образов m не должно превышать величины, примерно равной 0.15•n. Кроме того, если два образа А и Б сильно похожи, они, возможно, будут вызывать у сети перекрестные ассоциации, то есть предъявление на входы сети вектора А приведет к появлению на ее выходах вектора Б и наоборот.

Сеть Хэмминга

Когда нет необходимости, чтобы сеть в явном виде выдавала образец, то есть достаточно, скажем, получать номер образца, ассоциативную память успешно реализует сеть Хэмминга. Данная сеть характеризуется, по сравнению с сетью Хопфилда, меньшими затратами на память и объемом вычислений

Структура сети Хэмминга



– число образцов. Нейроны второго слоя связаны между собой ингибиторными синаптическими связями. Единственный синапс с положительной обратной связью для каждого нейрона соединен с его же аксоном.

Принцип работы

Идея работы сети состоит в нахождении расстояния Хэмминга от тестируемого образа до всех образцов.

Определение: Расстоянием Хэмминга называется число отличающихся битов в двух бинарных векторах.

Сеть должна выбрать образец с минимальным расстоянием Хэмминга до неизвестного входного сигнала, в результате чего будет активизирован только один выход сети, соответствующий этому образцу.

На стадии инициализации весовым коэффициентам первого слоя и порогу активационной функции присваиваются следующие значения:

$w_{ik} = \frac{x_i^k}{2}$ **, i=0...n-1, k=0...m-1, x_i^k – i-ый элемент k-ого образца**

$T_k = n / 2$, k = 0...m-1

Алгоритм работы

Весовые коэффициенты тормозящих синапсов во втором слое берут равными некоторой величине $0 < \varepsilon < 1/m$. Синапс нейрона, связанный с его же аксоном имеет вес +1.

2. На входы сети подается неизвестный вектор $X = \{x_i : i=0...n-1\}$, исходя из которого рассчитываются состояния нейронов первого слоя:

$$y_j^{(1)} = s_j^{(1)} = \sum_{i=0}^{n-1} w_{ij} x_i + T_j \quad , j = 0...m-1$$

После этого полученными значениями инициализируются значения аксонов второго слоя:

$$y_j^{(2)} = y_j^{(1)}, \; j = 0...m-1$$

6. Вычислить новые состояния нейронов второго слоя:

$$s_j^{(2)}(p+1) = y_j(p) - \varepsilon \sum_{k=0}^{m-1} y_k^{(2)}(p), \; k \neq j, \; j = 0...m-1$$

Алгоритм работы

1. **Вычислить новые значения аксонов:**

$$y_j^{(2)}(p+1) = f\left[s_j^{(2)}(p+1)\right], j = 0...m-1$$

Активационная функция f имеет вид порога, причем величина F должна быть достаточно большой, чтобы любые возможные значения аргумента не приводили к насыщению.

2. **Проверить, изменились ли выходы нейронов второго слоя за последнюю итерацию. Если да – перейди к шагу 2. Иначе – конец.**

Из оценки алгоритма видно, что роль первого слоя весьма условна: воспользовавшись один раз на шаге 1 значениями его весовых коэффициентов, сеть больше не обращается к нему, поэтому первый слой может быть вообще исключен из сети (заменен на матрицу весовых коэффициентов).

Методы и алгоритмы анализа структуры многомерных данных

Деменков П.С.,
Иванисенко В.А.

Кластерный анализ предназначен для разбиения множества объектов на заданное или неизвестное число классов на основании некоторого математического критерия качества классификации

(cluster (англ.) — гроздь, пучок, скопление, группа элементов, характеризуемых каким-либо общим свойством).

Критерий качества кластеризации

Критерий качества кластеризации в той или иной мере отражает следующие неформальные требования:

а) внутри групп объекты должны быть тесно связаны между собой;

б) объекты разных групп должны быть далеки друг от друга;

в) при прочих равных условиях распределения объектов по группам должны быть равномерными.

Требования а) и б) выражают стандартную концепцию компактности классов разбиения;

Требование в) состоит в том, чтобы критерий не навязывал объединения отдельных групп объектов.

Мера близости объектов

Узловым моментом в кластерном анализе считается выбор метрики (или меры близости объектов), от которого решающим образом зависит окончательный вариант разбиения объектов на группы при заданном алгоритме разбиения. В каждой конкретной задаче этот выбор производится по-своему, с учетом главных целей исследования, физической и статистической природы используемой информации и т. п. При применении экстенсиональных методов распознавания, как было показано в предыдущих лекциях, выбор метрики достигается с помощью специальных алгоритмов преобразования исходного пространства признаков.

Расстояние между группами объектов

Другой важной величиной в кластерном анализе является расстояние между целыми группами объектов. Приведем примеры наиболее распространенных расстояний и мер близости, характеризующих взаимное расположение отдельных групп объектов.

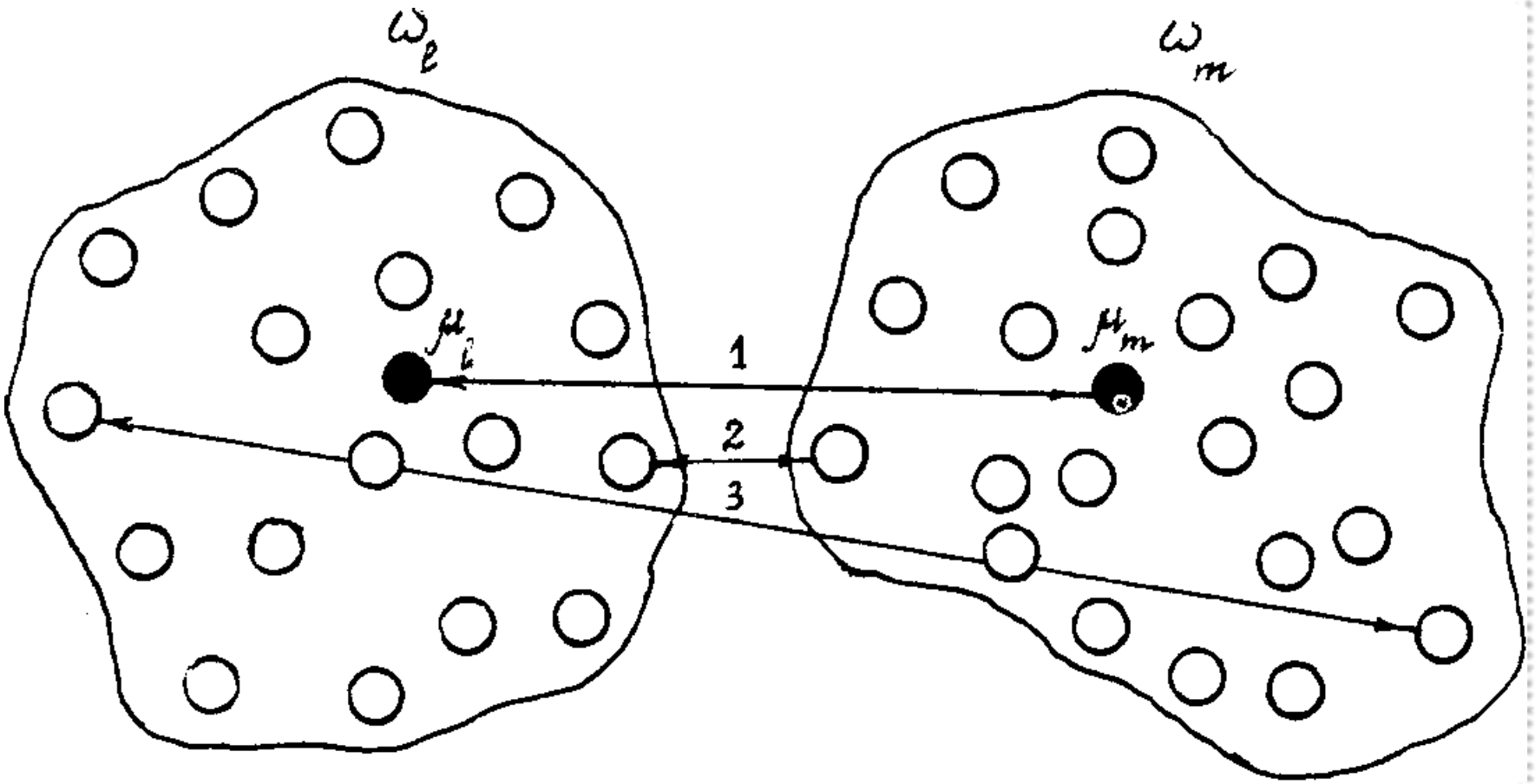
Пусть w_i — i-я группа (класс, кластер) объектов,

N_i — число объектов, образующих группу w_i,

вектор μ_i — среднее арифме-тическое объектов, входящих в w_i (другими словами [μ_i — «центр тяжести» i-й группы),

а $q(w_l, w_m)$ — расстояние между группами w_l и w_m

Различные способы определения расстояния между кластерами



Формулы для вычисления расстояния

Расстояние ближайшего соседа есть расстояние между ближайшими объектами кластеров:

$$q_{\min}(w_l, w_m) = \min_{x_i \in w_l, x_j \in w_m} d(x_i, x_j)$$

Расстояние дальнего соседа — расстояние между самыми дальними объектами кластеров:

$$q_{\max}(w_l, w_m) = \max_{x_i \in w_l, x_j \in w_m} d(x_i, x_j)$$

Расстояние центров тяжести равно расстоянию между центральными точками кластеров:

$$q(w_1, w_m) = d(\mu_1, \mu_m)$$

Формулы для вычисления расстояния

Обобщенное (по Колмогорову) расстояние между классами, или обобщенное K-расстояние, вычисляется по формуле

$$q_{\tau}^{(K)}(w_l, w_m) = \left[\frac{1}{N_l N_m} \sum_{x_i \in w_l} \sum_{x_j \in w_m} d^{\tau}(x_i, x_j) \right]^{\frac{1}{\tau}}$$

В частности, при $\tau \to \infty$ и при $\tau \to -\infty$ имеем

$$q_{\infty}^{(K)}(w_1, w_m) = q_{max}(w_1, w_m)$$

$$q_{-\infty}^{(K)}(w_1, w_m) = q_{min}(w_1, w_m)$$

Влияние меры на получаемый результат

Выбор той или иной меры расстояния между кластерами влияет, главным образом, на вид выделяемых алгоритмами кластерного анализа геометрических группировок объектов в пространстве признаков. Так, алгоритмы, основанные на расстоянии ближайшего соседа, хорошо работают в случае группировок, имеющих сложную, в частности, цепочечную структуру. Расстояние дальнего соседа применяется, когда искомые группировки образуют в пространстве признаков шаровидные облака. И промежуточное место занимают алгоритмы, использующие расстояния центров тяжести и средней связи, которые лучше всего работают в случае группировок эллипсоидной формы.

Последствия выбора не правильной меры

Нацеленность алгоритмов кластерного анализа на определенную структуру группировок объектов в пространстве признаков может приводить к неоптимальным или даже неправильным результатам, если гипотеза о типе группировок неверна. В случае отличия реальных распределений от гипотетических указанные алгоритмы часто «навязывают» данным не присущую им структуру и дезориентируют исследователя. Поэтому экспериментатор, учитывающий данный факт, в условиях априорной неопределенности прибегает к применению батареи алгоритмов кластерного анализа и отдает предпочтение какому-либо выводу на основании комплексной оценки совокупности результатов работы этих алгоритмов.

Критерии качества разбиения

Многообразие алгоритмов кластерного анализа обусловлено также множеством различных критериев, выражающих те или иные аспекты качества автоматического группирования. Простейший критерий качества непосредственно базируется на величине расстояния между кластерами. Однако такой критерий не учитывает «населенность» кластеров — относительную плотность распределения объектов внутри выделяемых группировок. Поэтому другие критерии основываются на вычислении средних расстояний между объектами внутри кластеров. Но наиболее часто применяются критерии в виде отношений показателей «населенности» кластеров к расстоянию между ними. Это, например, может быть отношение суммы межклассовых расстояний к сумме внутриклассовых (между объектами) расстояний или отношение общей дисперсии данных к сумме внутриклассовых дисперсий и дисперсии центров кластеров.

Функционалы качества и конкретные алгоритмы автоматической классификации достаточно полно и подробно рассмотрены в специальной литературе. Эти функционалы и алгоритмы характеризуются различной трудоемкостью и подчас требуют ресурсов высокопроизводительных компьютеров. Разнообразные процедуры кластерного анализа входят в состав практически всех современных пакетов прикладных программ для статистической обработки многомерных данных.

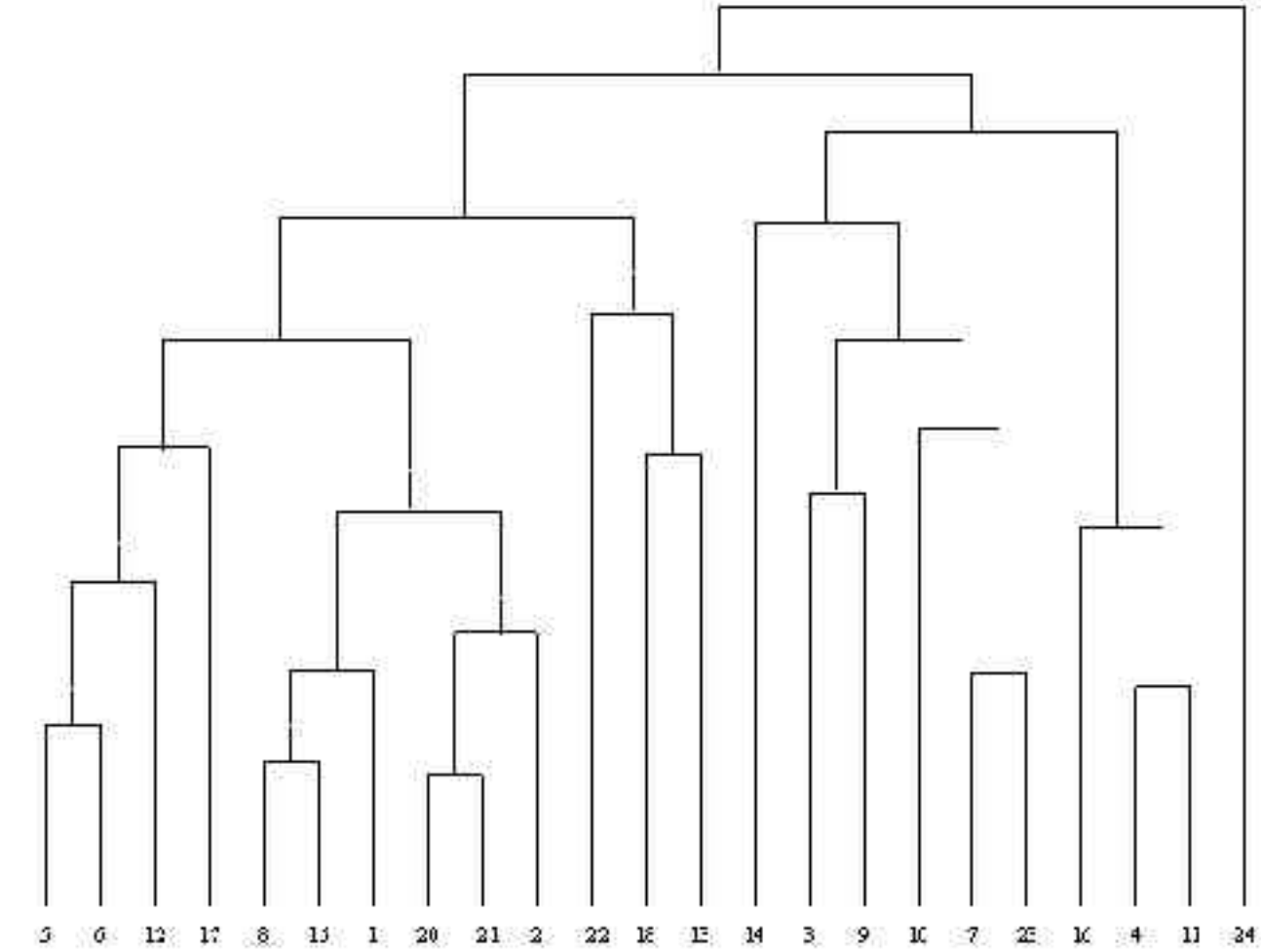
Иерархическое группирование

Классификационные процедуры иерархического типа предназначены для получения наглядного представления о стратификационной структуре всей исследуемой совокупности объектов. Эти процедуры основаны на последовательном объединении кластеров (агломеративные процедуры) и на последовательном разбиении (дивизимные процедуры). Наибольшее распространение получили агломеративные процедуры.

Агломеративные процедуры

На первом шаге все объекты считаются отдельными кластерами. Затем на каждом последующем шаге два ближайших кластера объединяются в один. Каждое объединение уменьшает число кластеров на один так, что в конце концов все объекты объединяются в один кластер. Наиболее подходящее разбиение выбирает чаще всего сам исследователь, которому предоставляется дендрограмма, отображающая результаты группирования объектов на всех шагах алгоритма. Могут одновременно также использоваться и математические критерии качества группирования.

Результаты работы иерархической агломеративной процедуры группирования объектов, представленные в виде дендрограммы.



Принцип определения расстояния

Различные варианты определения расстояния между кластерами дают различные варианты иерархических агломеративных процедур. Учитывая специфику подобных процедур, для задания расстояния между классами оказывается достаточным указать порядок пересчета расстояний между классом w_l и классом $w(m, n)$ являющимся объединением двух других классов w_m и w_n по расстояниям $q_{mn} = q(w_m, w_n)$ и $q_{ln} = q(w_l, w_n)$ между этими классами.

Формула для вычисления расстояния

В литературе предлагается следующая общая формула для вычисления расстояния между некоторым классом w_l и классом $w(m, n)$:

$$q_{l(m,n)} = q(w_l, w(m, n)) = \alpha q_{lm} + \beta q_{ln} + \gamma q_{mn} + \delta \mid q_{lm} - q_{ln} \mid,$$

где α, β, γ и δ — числовые коэффициенты, определяющие нацеленность агломеративной процедуры на решение той или иной экстремальной задачи. В частности, полагая $\alpha = \beta = -\delta = ½$ и $\gamma = 0$, приходим к расстоянию, измеряемому по принципу ближайшего соседа. Если положить $\alpha = \beta = \delta = ½$ и $\gamma = 0$, то расстояние между двумя классами определится как расстояние между двумя самыми далекими объектами этих классов, то есть это будет расстояние дальнего соседа.

Варианты определения коэффициентов

выбор коэффициентов соотношения по формулам

$$\alpha = \frac{N_m}{N_m + N_n}, \quad \beta = \frac{N_n}{N_m + N_n}, \quad \gamma = \delta = 0$$

приводит к расстоянию $q_{ср}$ между классами, вычисленному как среднее расстояние между всеми парами объектов, один из которых берется из одного класса, а другой из другого.

Использование следующей модификации формулы

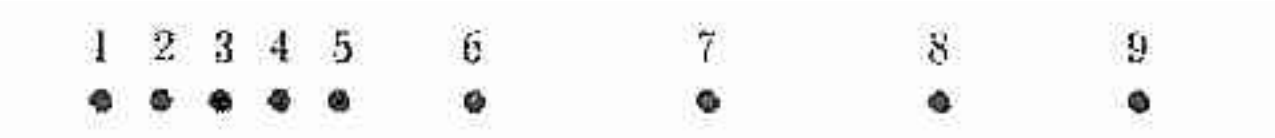
$$q^2_{l(m,n)} = \frac{N_l + N_m}{N_l + N_m + N_n} q^2_{lm} + \frac{N_l + N_n}{N_l + N_m + N_n} q^2_{ln} - \frac{N_l}{N_l + N_m + N_n} q^2_{mn}$$

дает агломеративный алгоритм, приводящий к минимальному увеличению общей суммы квадратов расстояний между объектами внутри классов на каждом шаге объединения этих классов.

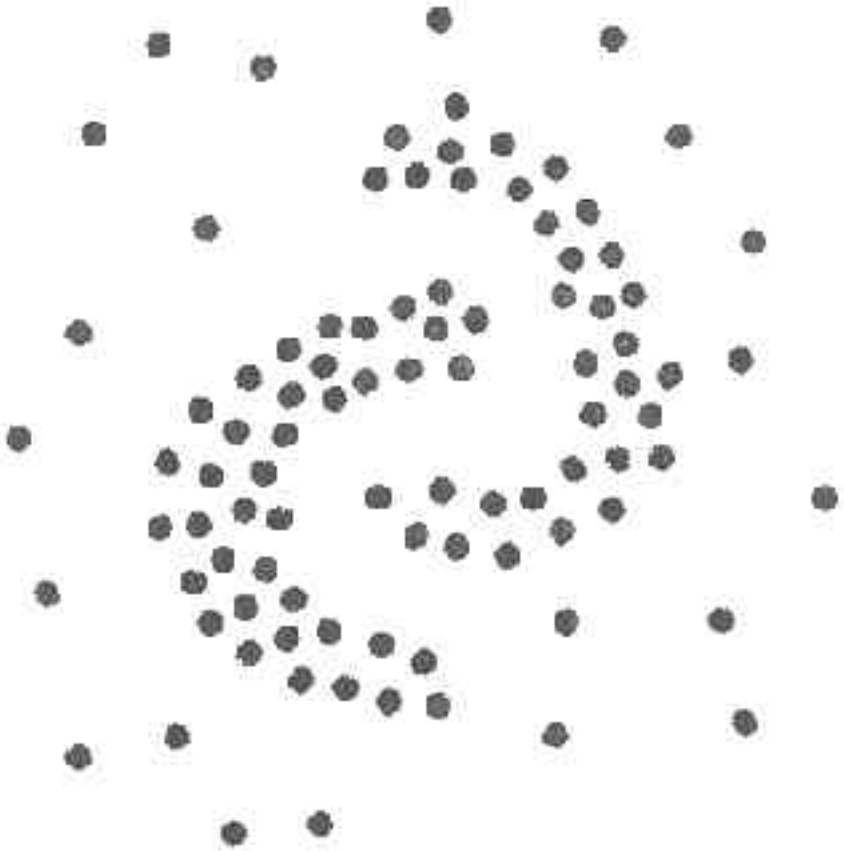
В отличие от оптимизационных кластерных алгоритмов предоставляющих исследователю конечный результат группирования объектов, иерархические процедуры позволяют проследить процесс выделения группировок и иллюстрируют соподчиненность кластеров, образующихся на разных шагах какого-либо агломеративного или дивизимного алгоритма.

Гипотеза λ компактности

Гипотеза компактности оперирует абсолютными значениями расстояний между векторами в пространстве характеристик. Однако на некоторых примерах можно показать, что важную роль в задачах анализа данных играют не только сами расстояния, но и отношения между ними.



Пример



Зрительный аппарат человека обладает уникальными способностями делать классификацию (таксономию) множества объектов, если они представлены точками на плоскости

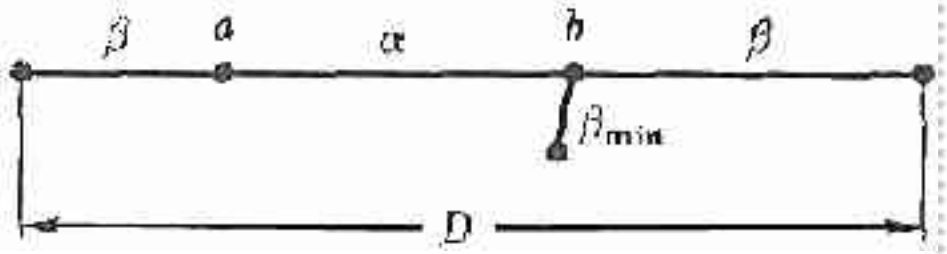
λ расстояние

Формулировка гипотезы λ-компактности опирается на понятие λ -расстояния, которое учитывает нормированное расстояние d между элементами множества и характеристику τ локальной плотности множества в окрестностях этих элементов.

Алгоритм расчета λ расстояния

1. Вычислить диаметр графа. ($D = \max \rho(a_i, a_j)$)
2. Нормированное расстояние $d = \alpha / D$
3. Найти самое короткое ребро, смежное с ребром (ab). Его длину обозначим через β_{min}
4. Отношение длин этих смежных отрезков обозначим через $\tau^* = \alpha / \beta_{min}$
5. Нормируем τ^*. Величина $\tau = \tau^* / \tau_{max}$ является нормированной характеристикой локальной неоднородности плотности множества в окрестностях точек a и b.

Величину $\lambda = f(\tau, d)$ называем λ-расстоянием между точками a и b



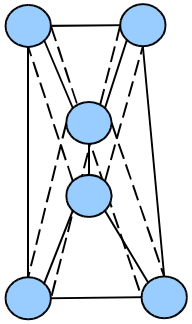
Кратчайший незамкнутый путь

Кратчайшим незамкнутым путём (КНП) называется граф, соединяющий между собой все вершины, не имеет петель и суммарная длина ребер минимальна.

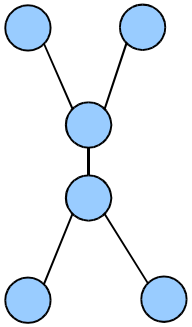
Алгоритм построения:

3. **Соединяются ребром самые близкие точки**
4. **Для пары самых близких точек проверяется нет ли уже построенного пути между ними, если нет, то соединяются ребром.**

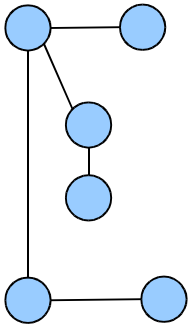
Примеры



Полный граф



КНП



λ-КНП

Критерий «равномощности» классов

При «ручной» таксономии человек стремится к такому решению, при котором граница между таксонами проходила бы по участку с наибольшим значением характеристики $\lambda = f(\tau, d)$, которую мы называем λ-расстоянием. Если имеется несколько возможных вариантов кластеризации с примерно одинаковыми значениями λ, то предпочтение отдается тому варианту, при котором классы включают в свой состав по возможности одинаковое количество объектов. Этот критерий «равномощности» классов хорошо отражает величина

$$h = k^k \prod_{i=1}^{k} \frac{m_i}{m}$$

где k — количество таксонов, m_i — число объектов в i-м таксоне, а m — общее число объектов.

Оценка качества разбиения на классы

Характеристикой качества кластеризации, является величина

$$F = F(h,\tau,d)$$

Исследования, связанные с аккуратной формулировкой гипотезы λ-компактности, привели к заключению, что человек в процессе таксономии действительно использует все эти составляющие **h**, **d**, **τ**, но придает им разный вес.

Была сформулирована задача идентификации модели следующего вида:

$$F = h^q \tau^s d^v.$$

наибольший вес придается нормированному расстоянию **d**, затем характеристике скачка плотности **τ** и лишь потом характеристике равномощности таксонов **h**

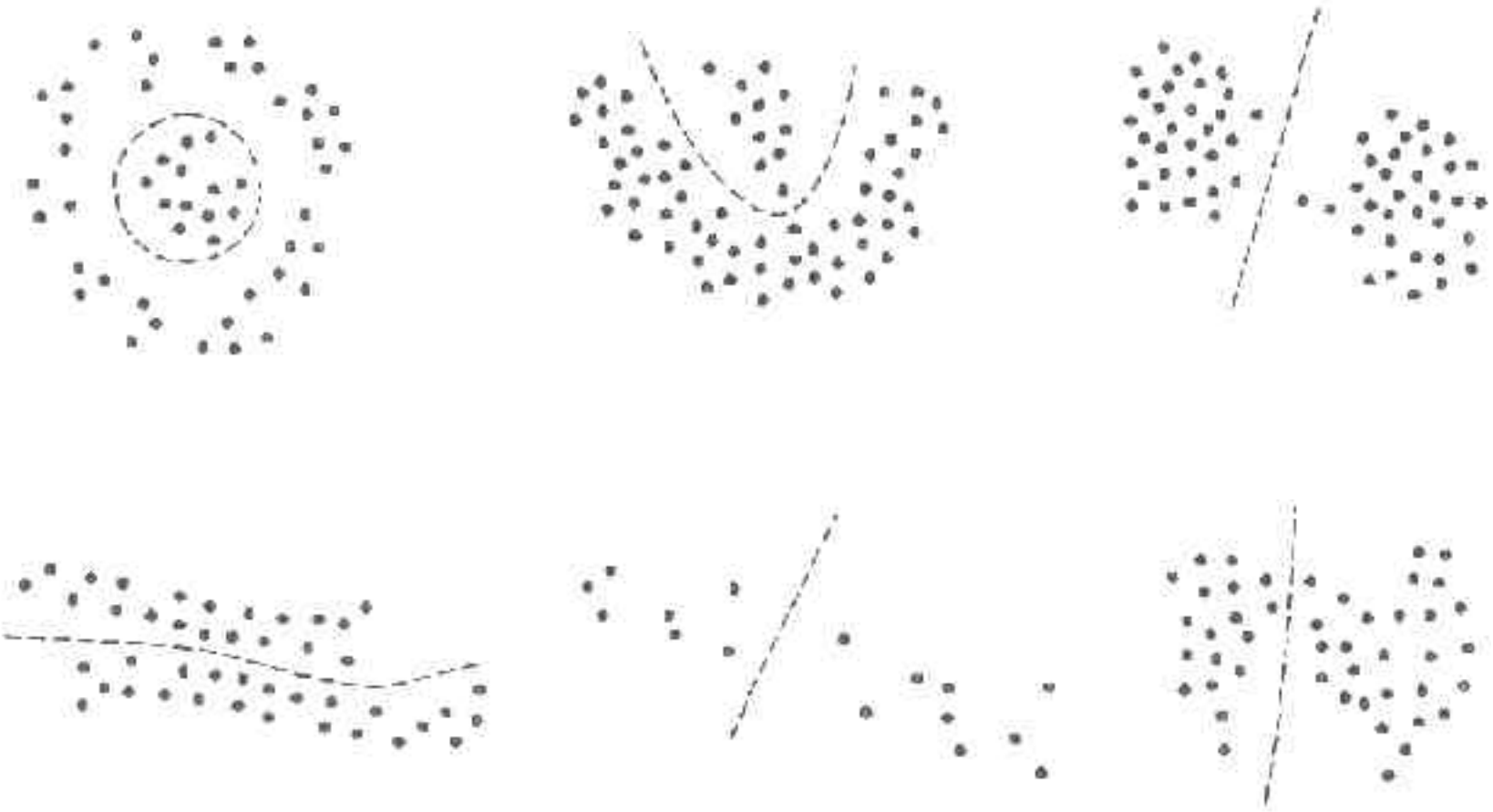
$$F = h^4 \tau^2 d^1.$$

Описание алгоритма λ-КРАБ

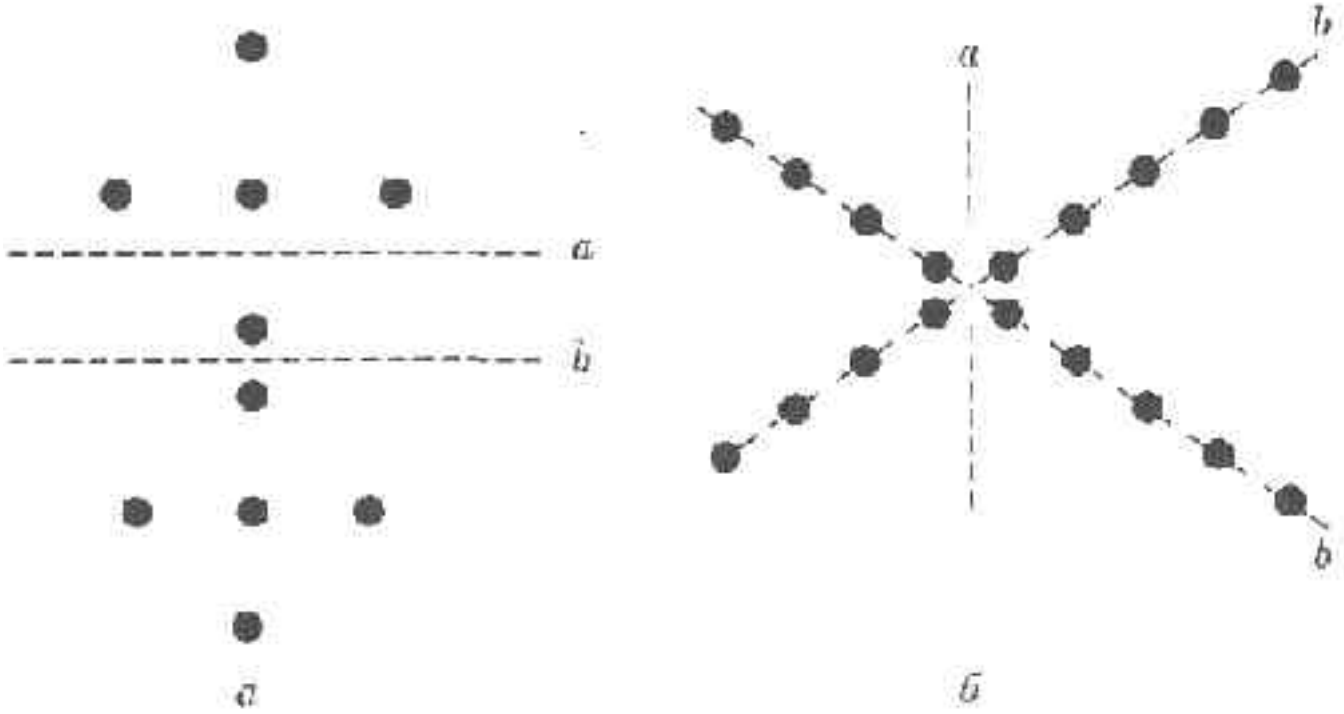
1. Строится λ-КНП
2. Для разбиения множества A на два таксона необходимо разорвать одно ребро из ребер графа λ-КНП. Выберем ребро j с λ-длиной $\lambda_j = \tau_j^2 d_j$. Оставшимися ребрами λ-КНП соединяются два подмножества по m_i точек в каждом i-м подмножестве
3. Вычисляется характеристику равномощности таксонов h_j

Общая оценка качества F_j этого j-го варианта таксономии равна $\lambda_j h_j^4$. Вычисление величины F_j для всех $(m — 1)$ ребер графа позволяет найти такой вариант разбиения, при котором достигается максимум критерия F.

Примеры



Примеры не совпадения кластеризации



Выбор числа классов

Если желательное число классов задано диапазоном от k_{min} До k_{max}, то, наблюдая за функцией F = f(k), можно в заданных пределах найти число таксонов, при котором F достигает максимума, что соответствует наиболее предпочтительной способу кластеризации.

Выбор числа классов