Лекция 10

Способы представления исходной информации в интеллектуальных системах

Человек, решающий задачу выбора целесообразного поведения в той или иной ситуации, прежде всего анализирует существенные и несущественные обстоятельства, влияющие на принимаемое решение. Процесс выделения существенных для данной задачи обстоятельств можно представить как разбиение входных ситуаций на классы, обладающие тем свойством, что все ситуации из одного класса требуют одних и тех же действий.

Оценка входной ситуации человеком происходит на основе совокупности сигналов, поступающих от его органов чувств. На основании этих сигналов мозг вырабатывает команды, которые обеспечивают реакцию человека на ситуацию. Сигналы поступают от рецепторов (зрительных, тактильных и др.). Совокупность таких сигналов формирует представление человека о ситуации.

Вычислительная машина, на которой моделируется аналогичный процесс, должна обладать возможностью получать описание входной ситуации от внешних «рецепторов» в виде различных наборов данных. Очевидно, объем информации, который получает компьютер, несоизмеримо меньше объемов информации, с которыми имеет дело человек; кроме того, такая информация будет представлена исключительно в численной форме.

Для того, чтобы эффективно оценить, относятся ли различные ситуации к одному классу, интеллектуальная система должна иметь возможность рассмотреть и оценить ряд конкретных примеров таких ситуаций, включенных в обучающее множество.

Обучение на основе примеров является типичным случаем индуктивного обучения и широко используется в интеллектуальных системах. На основе предъявленных примеров (и, возможно, контрпримеров) интеллектуальная система должна сформировать общее понятие, охватывающее примеры и исключающее контрпримеры.

Источником примеров, на которых осуществляется обучение, может быть учитель то есть лицо, которое заранее знает концепцию формируемого понятия и подбирает наиболее удачные обучающие выборки.

Источником примеров для обучения может быть внешняя среда, с которой взаимодействует интеллектуальная система. В этом случае обучающие выборки формируются случайным образом в зависимости от внешних факторов. Обучение на таких выборках существенно сложнее.

Наконец, источником примеров для обучения может стать сама интеллектуальная система. Например, в случае взаимодействия интеллектуального робота с внешней средой действия самого робота могут привести к созданию обучающей выборки, то есть образуется множество сходных ситуаций с известными результатами, которые можно затем обобщить.

Для системы машинного обучения принципиально важным является вопрос, что поступает на вход системы, в каком виде предъявляются примеры понятия, включенные в состав обучающего множества.

Все основные методы решения задач индуктивного построения понятий базируются на концепции признакового описания примера понятия, а именно: любой элемент обучающей выборки, который может быть представлен в системе, полностью определяется набором свойств, или признаков. Такое задание объекта исследования называется признаковым описанием объекта.

Значения, которые могут принимать признаки объекта, относятся к трем основным типам: количественные или числовые, качественные и шкалированные.

В случае числовых признаков на множестве значений признаков может быть введена метрика, позволяющая дать количественную оценку значения признака. Часто такие значения являются результатом измерений физических величин, таких, как длина, вес, температура и др.

В случае, если признаки могут иметь качественный характер, но при этом их значения можно упорядочить друг относительно друга, говорят, что такие значения образуют ранговую или порядковую шкалу.

Примерами таких шкал порядка могут быть ряды типа {большой, средний, маленький} или {горячий, теплый, холодный}. С помощью таких шкал порядка можно судить, какой из двух объектов является наилучшим, но нельзя оценить, сколь близки или далеки эти объекты по некоторому критерию.

Третий случай заключается в том, что значения признаков имеют чисто качественный характер, связать эти значения между собой не удается. Примерами таких значений могут быть цвет = {красный, желтый, зеленый} или материал = {стекло, дерево, пластмасса, железо}.

Таблица 1. Выборка K^+

возраст	пол	стадия	Что болит	температура	Прочие ощущения
Пожилой	Жен.	Средняя	Голова	36,6	Озноб
Пожилой	Муж:.	Средняя	Голова	36,4	Озноб
Пожилой	Муж.	Поздняя	Голова	36,7	Озноб
Средний	Жен.	Начальная	Голова	37,8	Никаких
Пожилой	Жен.	Начальная	Нет	38,0	Нет аппетита
Пожилой	Муж:.	Начальная	Спина	36,3	Нет аппетита
Пожилой	Муж.	Начальная	Нет	36,6	Нет аппетита
Пожилой	Муж.	Начальная	Голова	38,2	Озноб

Выборка К-

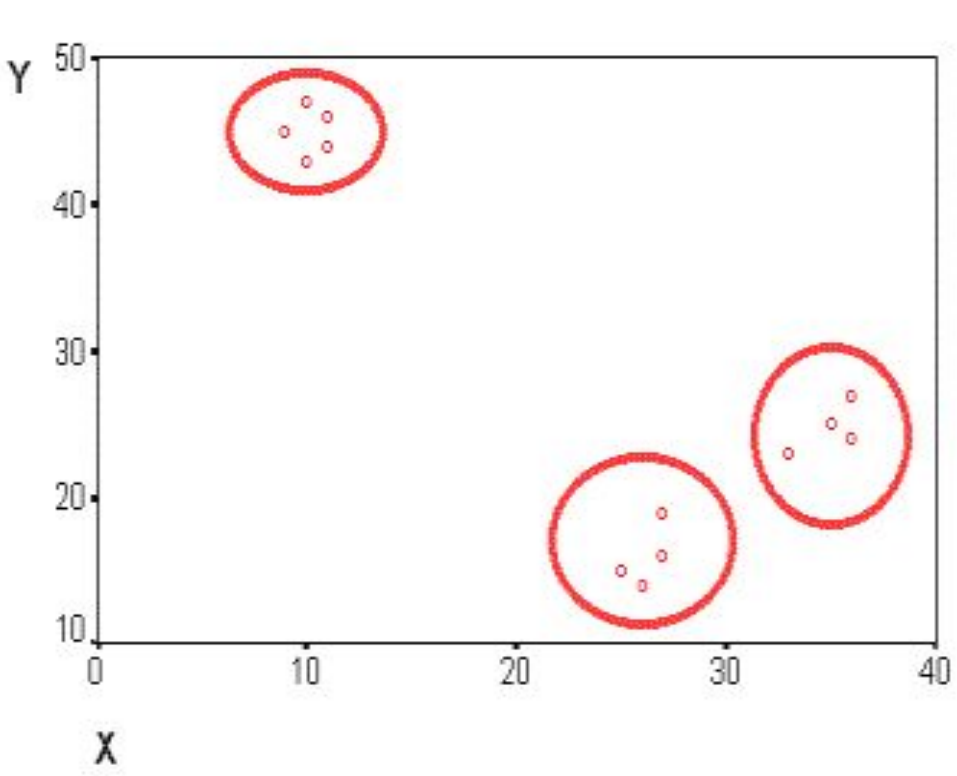
возраст	пол	стадия	Что болит	температура	Прочие ощущения
Молодой	Муж.	-	Нет	36,6	Никаких
Средний	Жен.	Средняя	Живот	37,7	Тошнота
Пожилой	Жен.	Начальная	Руки	36,7	Никаких
Пожилой	Муж.	Поздняя	Спина	36,3	Нет аппетита

- Термин кластерный анализ, впервые введенный Трионом (Tryon) в 1939 году, включает в себя более 100 различных алгоритмов.
- В отличие от задач классификации, кластерный анализ не требует априорных предположений о наборе данных, не накладывает ограничения на представление исследуемых объектов, позволяет анализировать показатели различных типов данных (интервальным данным, частотам, бинарным данным). При этом необходимо помнить, что переменные должны измеряться в сравнимых шкалах.

- Рассмотрим пример процедуры кластерного анализа.
- Допустим, мы имеем набор данных A, состоящий из 14-ти примеров, у которых имеется по два признака X и Y. Данные по ним приведены в таблице.

№ примера	признак X	признак Y
1	27	19
2	11	46
3	25	15
4	36	27
5	35	25
6	10	43
7	11	44
8	36	24
9	26	14
10	26	14
11	9	45
12	33	23
13	27	16
14	10	47

- Данные в табличной форме не носят информативный характер. Представим переменные X и Y в виде диаграммы рассеивания



- На рисунке мы видим несколько групп "похожих" примеров. Примеры (объекты), которые по значениям X и Y "похожи" друг на друга, принадлежат к одной группе (кластеру); объекты из разных кластеров не похожи друг на друга.
- Критерием для определения схожести и различия кластеров является расстояние между точками на диаграмме рассеивания. Это сходство можно "измерить", оно равно расстоянию между точками на графике. Способов определения меры расстояния между кластерами, называемой еще мерой близости, существует несколько.

- Наиболее распространенный способ - вычисление евклидова расстояния между двумя точками i и j на плоскости, когда известны их координаты X и Y:

$$D_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}$$

- Наиболее распространенный способ - вычисление евклидова расстояния между двумя точками i и j на плоскости, когда известны их координаты X и Y:

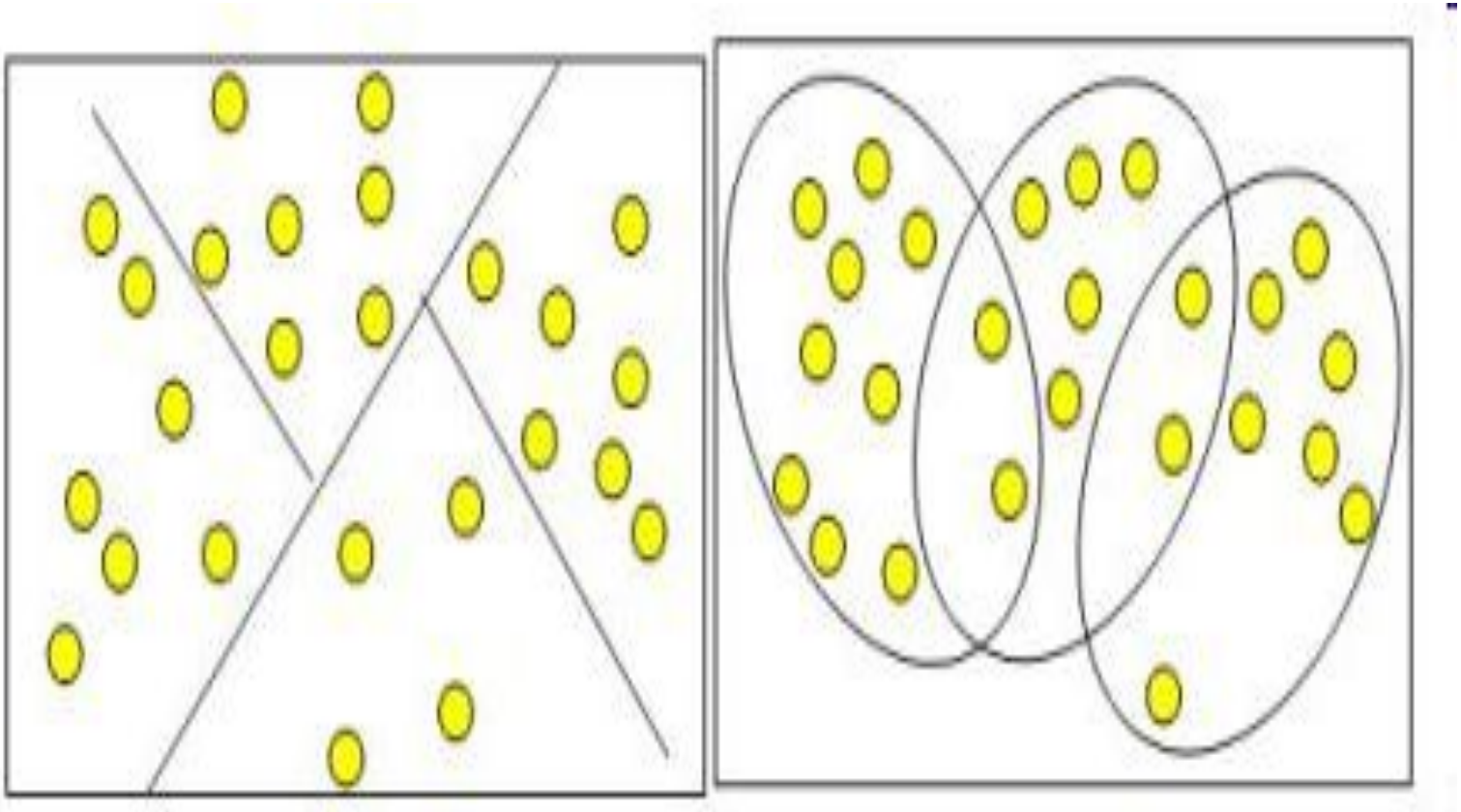
$$D_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}$$

Аналогично для трех координат

Кластер имеет следующие математические характеристики: центр, радиус, среднеквадратическое отклонение, размер кластера.

- Центр кластера - это среднее геометрическое место точек в пространстве переменных.
- Радиус кластера - максимальное расстояние точек от центра кластера.
- Кластеры могут быть перекрывающимися. Такая ситуация возникает, когда обнаруживается перекрытие кластеров. В этом случае невозможно при помощи математических процедур однозначно отнести объект к одному из двух кластеров. Такие объекты называют спорными.

- Спорный объект - это объект, который по мере сходства может быть отнесен к нескольким кластерам.
- Размер кластера может быть определен либо по радиусу кластера, либо по среднеквадратичному отклонению объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до центра кластера меньше радиуса кластера. Если это условие выполняется для двух и более кластеров, объект является спорным.
- Неоднозначность данной задачи может быть устранена экспертом или аналитиком.



- Работа кластерного анализа опирается на два предположения. Первое предположение - рассматриваемые признаки объекта в принципе допускают желательное разбиение пула (совокупности) объектов на кластеры. В начале лекции мы уже упоминали о сравнимости шкал, это и есть второе предположение - правильность выбора масштаба или единиц измерения признаков.
- Выбор масштаба в кластерном анализе имеет большое значение.

- Рассмотрим пример. Представим себе, что данные признака x в наборе данных A на два порядка больше данных признака y: значения переменной x находятся в диапазоне от 100 до 700, а значения переменной y - в диапазоне от 0 до 1.
- Тогда, при расчете величины расстояния между точками, отражающими положение объектов в пространстве их свойств, переменная, имеющая большие значения, т.е. переменная x, будет практически полностью доминировать над переменной с малыми значениями, т.е. переменной y. Таким образом из-за неоднородности единиц измерения признаков становится невозможно корректно рассчитать расстояния между точками.

- Эта проблема решается при помощи предварительной стандартизации переменных. Стандартизация (standardization) или нормирование (normalization) приводит значения всех преобразованных переменных к единому диапазону значений путем выражения через отношение этих значений к некой величине, отражающей определенные свойства конкретного признака. Существуют различные способы нормирования исходных данных.

Наиболее распространенный:

- деление исходных данных на среднеквадратичное отклонение соответствующих переменных

- Наряду со стандартизацией переменных, существует вариант придания каждой из них определенного коэффициента важности, или веса, который бы отражал значимость соответствующей переменной. В качестве весов могут выступать экспертные оценки, полученные в ходе опроса экспертов - специалистов предметной области. Полученные произведения нормированных переменных на соответствующие веса позволяют получать расстояния между точками в многомерном пространстве с учетом неодинакового веса переменных

Методы кластерного анализа можно разделить на две группы:

- иерархические;
- неиерархические.

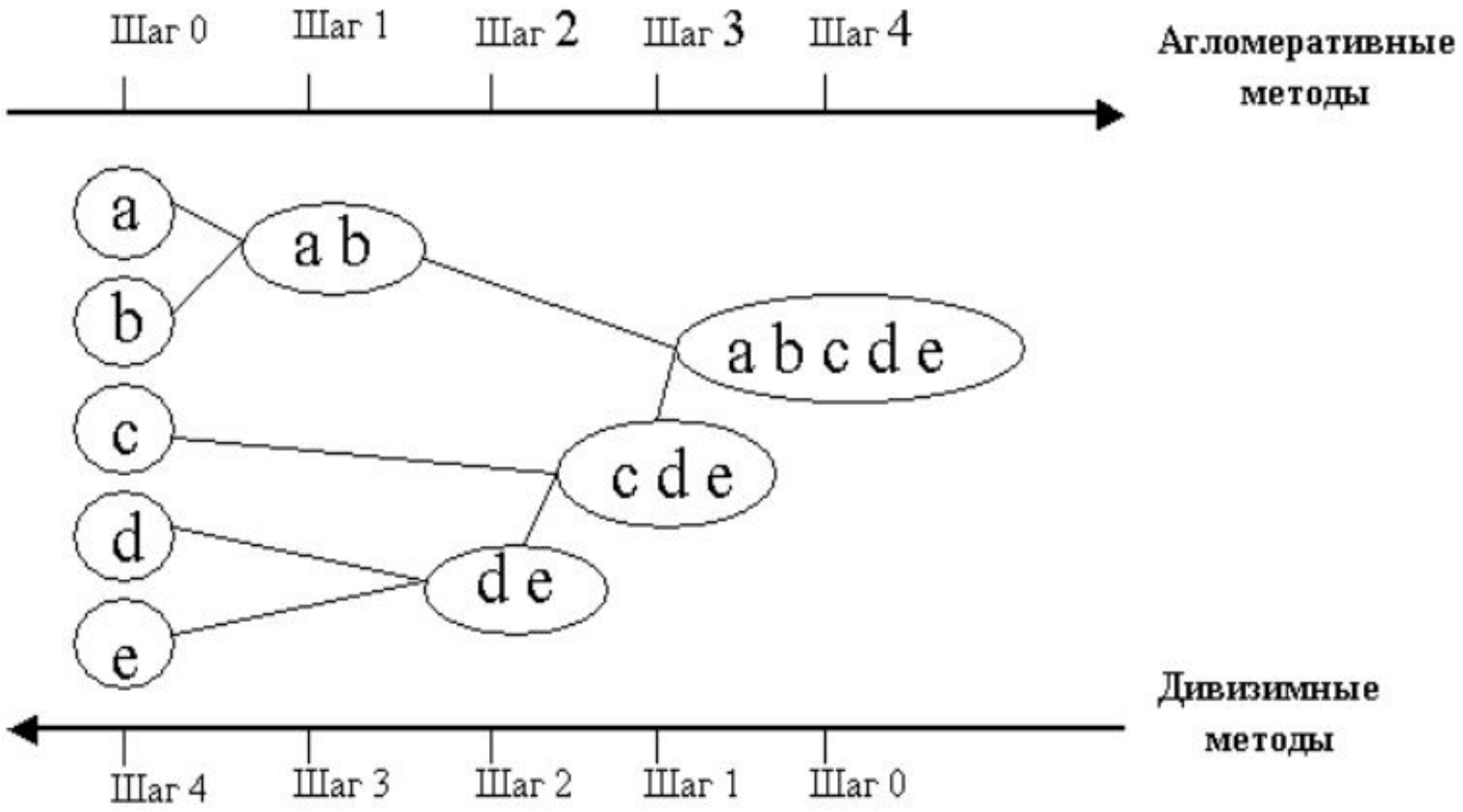
Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие

Иерархические агломеративные методы (Agglomerative Nesting, AGNES)

- Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.
- В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Иерархические дивизимные (делимые) методы (DIvisive ANAlysis, DIANA)

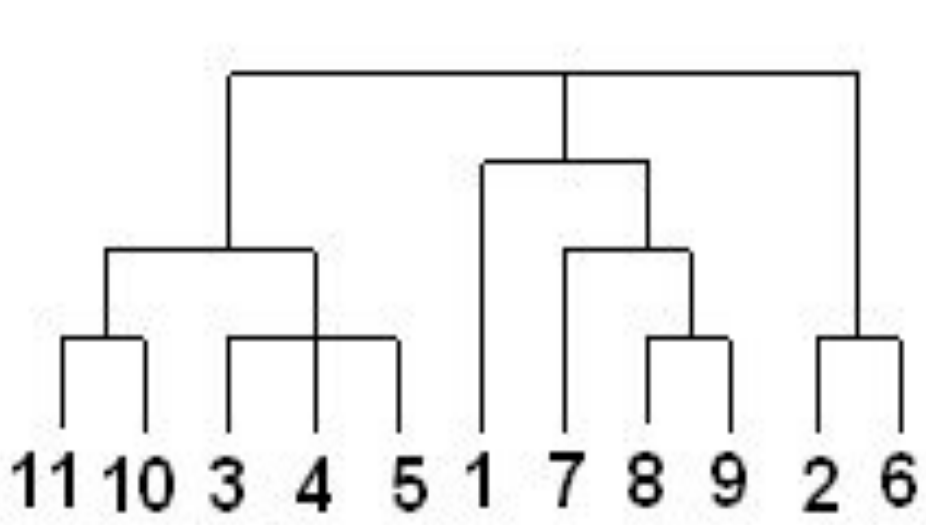
- Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.


Шаг 0
Шаг 1
Шаг 2
Шаг 3
Шаг 4
Агломеративные методы
a
b
c
d
e
a b
c d e
a b c d e
d e
Дивизимные методы
Шаг 4
Шаг 3
Шаг 2
Шаг 1
Шаг 0

- Программная реализация алгоритмов кластерного анализа широко представлена в различных инструментах Data Mining, которые позволяют решать задачи достаточно большой размерности. Например, агломеративные методы реализованы в пакете SPSS, дивизимные методы - в пакете Statgraf.
- Иерархические методы кластеризации различаются правилами построения кластеров. В качестве правил выступают критерии, которые используются при решении вопроса о "схожести" объектов при их объединении в группу (агломеративные методы) либо разделения на группы (дивизимные методы).
- Иерархические методы кластерного анализа используются при небольших объемах наборов данных.
- Преимуществом иерархических методов кластеризации является их наглядность.

- Иерархические алгоритмы связаны с построением дендрограмм (от греческого dendron - "дерево"), которые являются результатом иерархического кластерного анализа.
- Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.

- Существует много способов построения дендрограмм. В дендрограмме объекты могут располагаться вертикально или горизонтально. Пример вертикальной дендрограммы



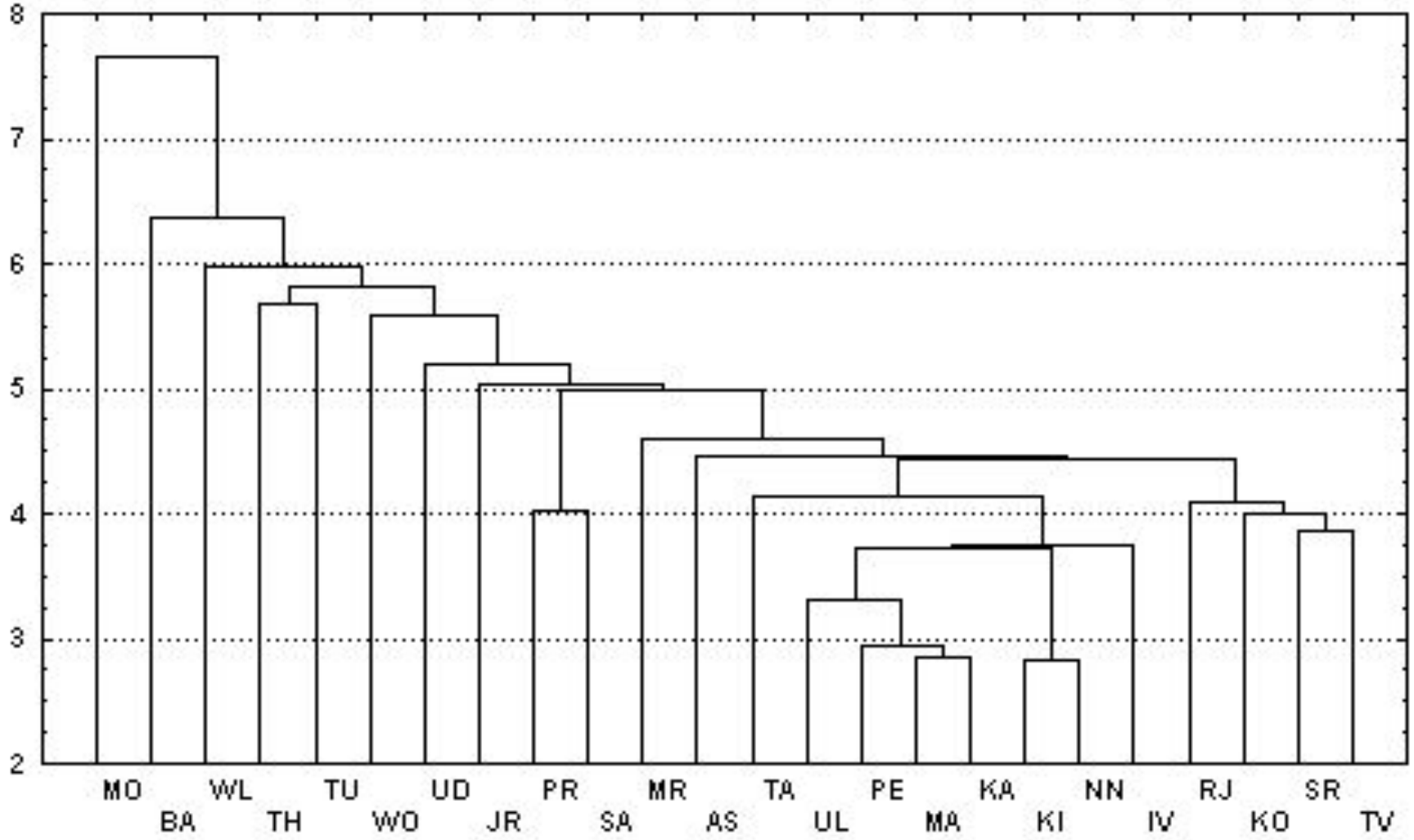
- Числа 11, 10, 3 и т.д. соответствуют номерам объектов или наблюдений исходной выборки. Мы видим, что на первом шаге каждое наблюдение представляет один кластер (вертикальная линия), на втором шаге наблюдаем объединение таких наблюдений: 11 и 10; 3, 4 и 5; 8 и 9; 2 и 6. На втором шаге продолжается объединение в кластеры: наблюдения 11, 10, 3, 4, 5 и 7, 8, 9. Данный процесс продолжается до тех пор, пока все наблюдения не объединятся в один кластер.

Методы объединения или связи

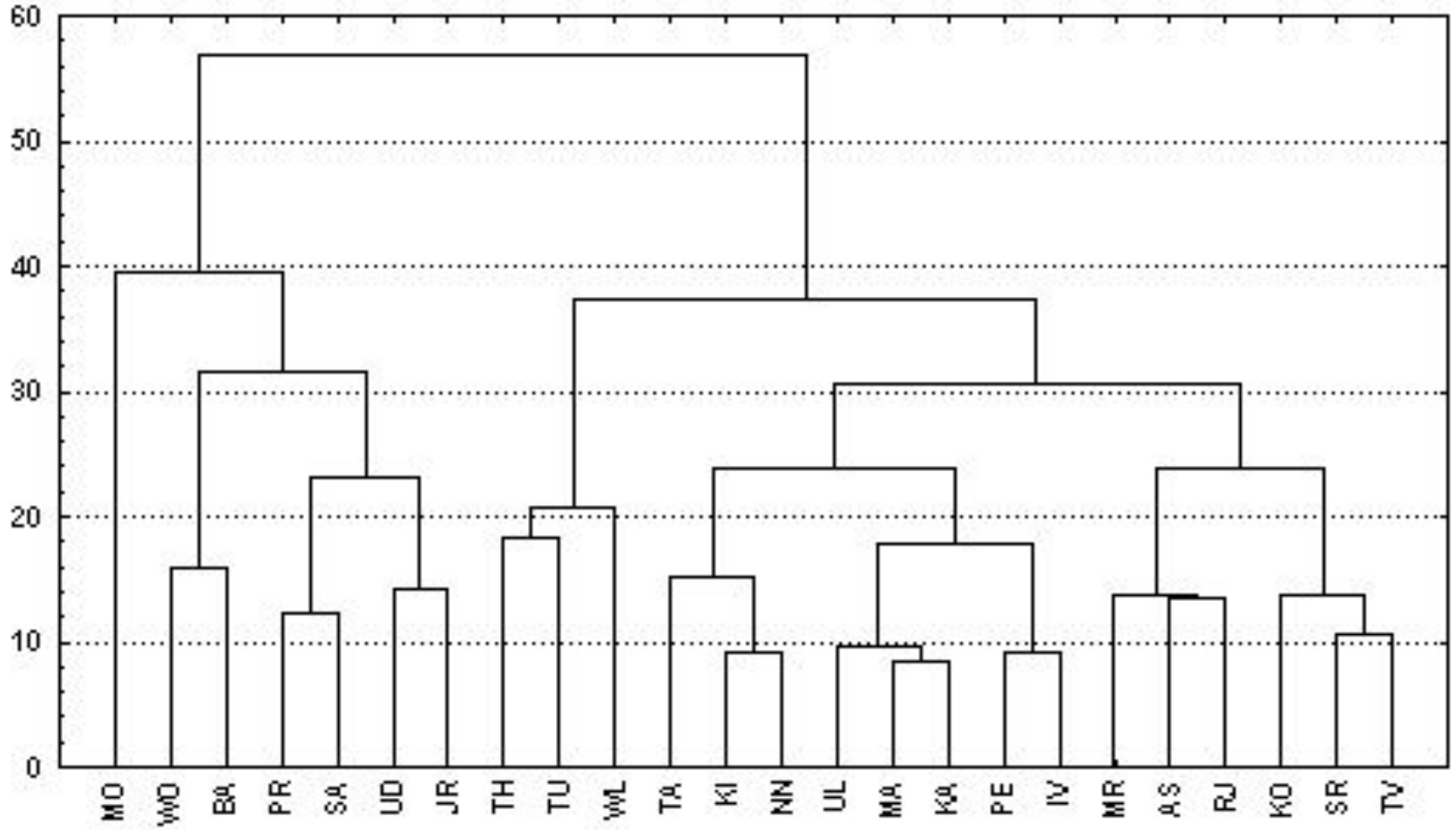
- Когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Возникает следующий вопрос - как определить расстояния между кластерами? Существуют различные правила, называемые методами объединения или связи для двух кластеров.

Метод ближнего соседа или одиночная связь. Здесь расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. Этот метод позволяет выделять кластеры сколь угодно сложной формы при условии, что различные части таких кластеров соединены цепочками близких друг к другу элементов.

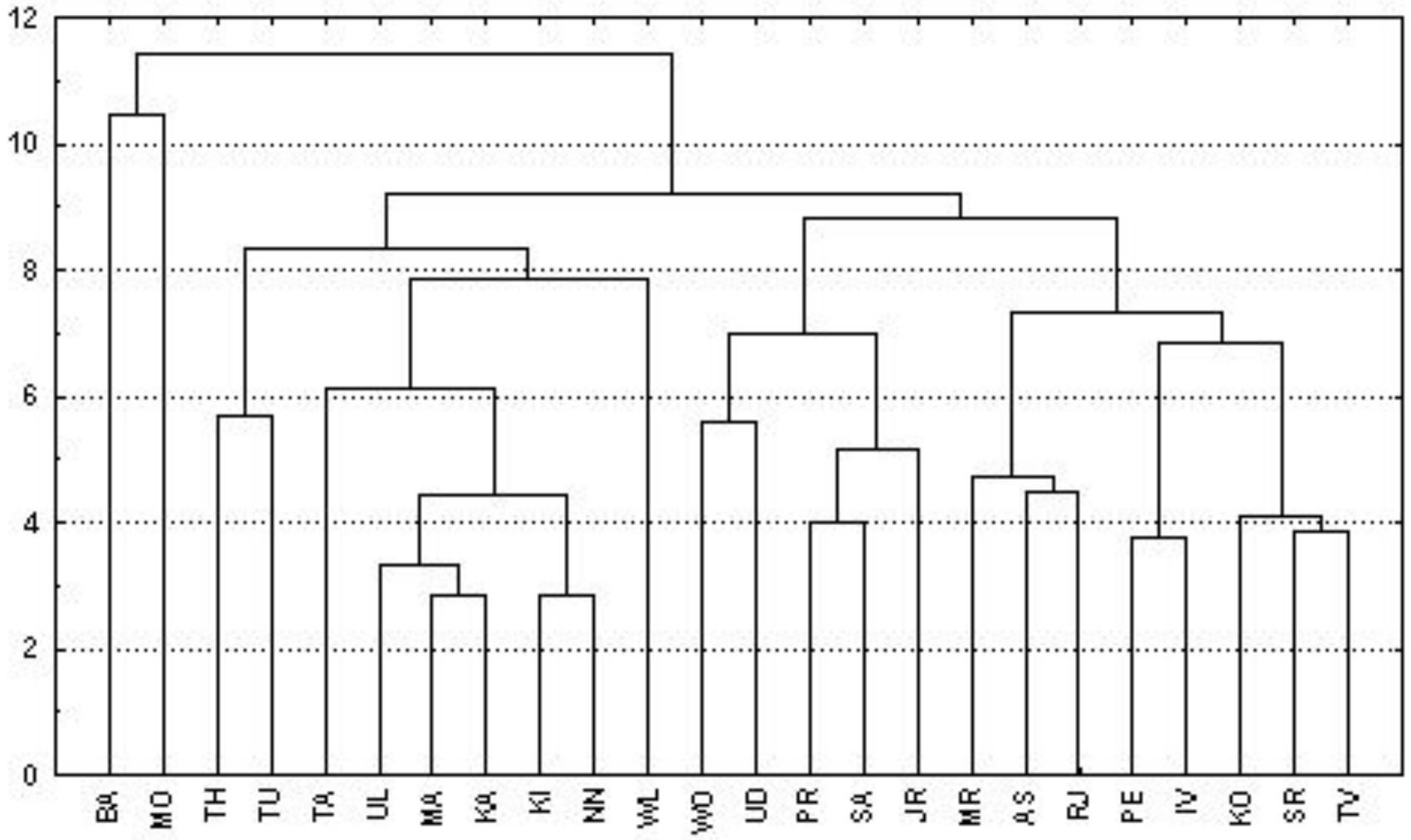
В результате работы этого метода кластеры представляются длинными "цепочками" или "волокнистыми" кластерами, "сцепленными вместе" только отдельными элементами, которые случайно оказались ближе остальных друг к другу.


8
7
6
5
4
3
2
MO
BA
WL
TH
TU
WO
UD
JR
PR
SA
MR
AS
TA
UL
PE
MA
KA
KI
NN
IV
RJ
KO
SR
TV

- **Метод Варда** (Ward's method). В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения (Ward, 1963). В отличие от других методов кластерного анализа для оценки расстояний между кластерами, здесь используются методы дисперсионного анализа. На каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров и "стремится" создавать кластеры малого размера.


60
50
40
30
20
10
0
MO
WO
BA
PR
SA
UD
JR
TH
TU
WL
TA
KI
NN
UL
MA
KA
PE
IV
MR
AS
RJ
KO
SR
TV

- **Метод наиболее удаленных соседей** или полная связь. Здесь расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями"). Метод хорошо использовать, когда объекты действительно происходят из различных "рощ". Если же кластеры имеют в некотором роде удлиненную форму или их естественный тип является "цепочечным", то этот метод не следует использовать.


12
10
8
6
4
2
0
BA
MO
TH
TU
TA
UL
MA
KA
KI
NN
WL
WO
UD
PR
SA
JR
MR
AS
RJ
PE
IV
KO
SR
TV

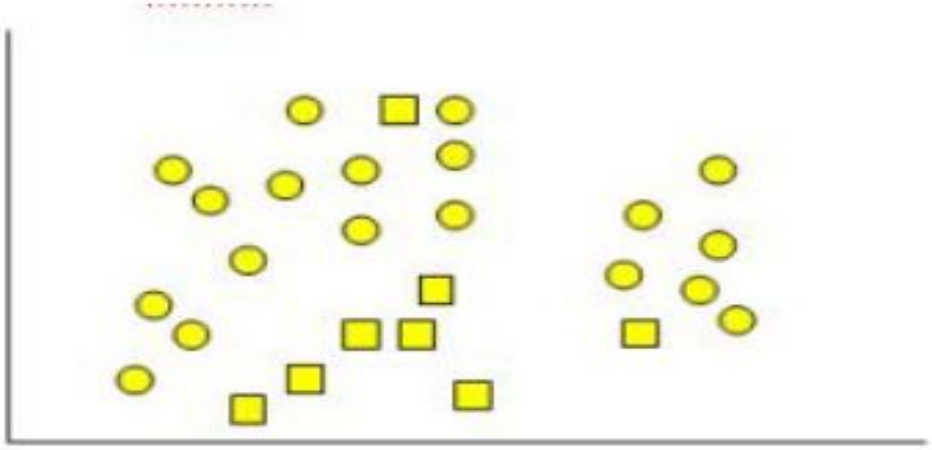
- **Метод невзвешенного попарного среднего** (метод невзвешенного попарного арифметического среднего - unweighted pair-group method using arithmetic averages, UPGMA (Sneath, Sokal, 1973)).
- В качестве расстояния между двумя кластерами берется среднее расстояние между всеми парами объектов в них. Этот метод следует использовать, если объекты действительно происходят из различных "рощ", в случаях присутствия кластеров "цепочного" типа, при предположении неравных размеров кластеров.

- **Метод взвешенного попарного среднего** (метод взвешенного попарного арифметического среднего - weighted pair-group method using arithmetic averages, WPGM A (Sneath, Sokal, 1973)). Этот метод похож на метод невзвешенного попарного среднего, разница состоит лишь в том, что здесь в качестве весового коэффициента используется размер кластера (число объектов, содержащихся в кластере).
- Этот метод рекомендуется использовать именно при наличии предположения о кластерах разных размеров.

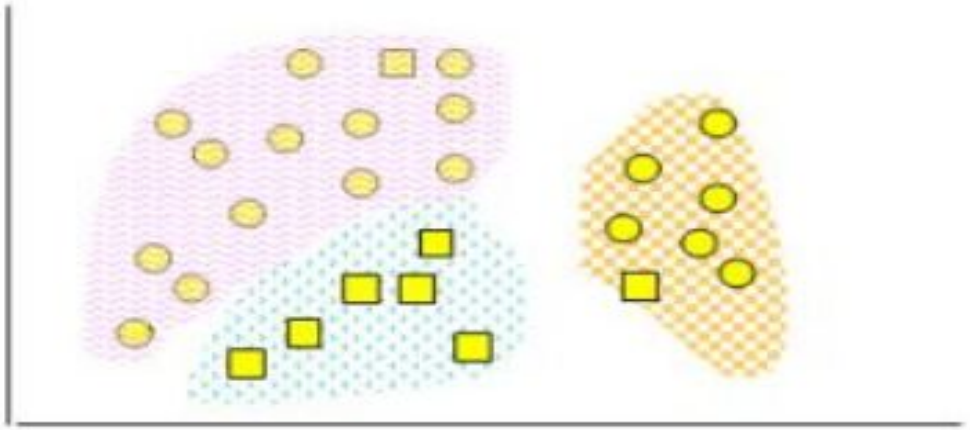
- **Невзвешенный центроидный метод** (метод невзвешенного попарного центроидного усреднения - unweighted pair-group method using the centroid average (Sneath and Sokal, 1973)).
- В качестве расстояния между двумя кластерами в этом методе берется расстояние между их центрами тяжести.

Таблица 5.2. Сравнение классификации и кластерзации

Характеристика	Классификация	Кластеризация
Контролируемость обучения	Контролируемое обучение	Неконтролируемое обучение
Стратегия	Обучение с учителем	Обучение без учителя
Наличие метки класса	Обучающее множество сопровождается меткой, указывающей класс, к которому относится наблюдение	Метки класса обучающего множества неизвестны
Основание для классификации	Новые данные классифицируются на основании обучающего множества	Дано множество данных с целью установления существования классов или кластеров данных



Классификация: классы предопределены изначально



Кластеризация: классы не предопределены, осуществляется поиск наиболее похожих, однородных групп

Важность алгоритмов “обучения без учителя” в том, что реальные признаки, описывающие объекты распознавания, очень часто бывают именно количественными, или числовыми. Известно, что человек плохо воспринимает информацию, представленную в виде больших наборов чисел. Первым и крайне важным этапом решения задачи обобщения в таком случае будет переход от количественных признаков к признакам качественным или хотя бы к шкалируемым. Здесь большую помощь могут оказать алгоритмы рассматриваемого типа.

Дадим более строгую формулировку задачи обучения «без учителя».

Пусть обучающая выборка содержит *M* объектов: $X = \{X_1, X_2, \ldots, X_n\}$- Каждый из этих объектов представляет собой n-мерный вектор X_i значений признаков:

$$X_i = < x_{i1}, x_{i2}, \ldots, x_{in} >,$$

где x_{ij} — значение *j-го* признака для i-го объекта, *n* — количество признаков, характеризующих объект.

Признаки, используемые для описания объекта, чисто количественные, к ним применимы введенные в предыдущей главе меры близости.

Требуется в соответствии с заданным критерием разделить набор *X* на классы, количество которых заранее неизвестно. Под критерием подразумевается мера близости всех объектов одного класса между собой. Будем считать, что работа алгоритма завершена успешно, если классы, сформированные в результате работы алгоритма, достаточно компактны и, возможно, выполнены некоторые дополнительные критерии.

При решении задачи обучения «без учителя» самыми несложными являются алгоритмы, основанные на мерах близости. Для достижения цели - компактного формирования классов — введем понятие точки-прототипа, или точки в *n*-мерном пространстве признаков, являющейся наиболее «типичной» представительницей построенного класса. В дальнейшем расстояние от объекта до класса будет заменяться расстоянием от объекта до точки-прототипа. Точка-прототип может быть сопоставлена каждому сформированному классу, и при этом вовсе не обязательно существование реального объекта, соответствующего точке-прототипу.

Алгоритм, основанный на понятии порогового расстояния

Пороговый алгоритм — один из самых несложных алгоритмов, базирующихся на понятии меры близости. Критерием отнесения объекта к классу здесь является пороговое расстояние *T.* Если объект находится в пределах порогового расстояния от точки-прототипа некоторого класса, то такой объект будет отнесен к данному классу. Если исследуемый объект находится на расстоянии, превышающем *T,* он становится прототипом нового класса.

Самая первая точка-прототип может выбираться произвольно. Результатом работы такого алгоритма будет разбиение объектов выборки *X* на классы, где в каждом классе расстояние между точкой-прототипом и любым другим элементом класса не превышает *T.* Пороговое расстояние *T* определим как половину расстояния между двумя наиболее удаленными друг от друга точками обучающей выборки.

Алгоритм

1. Выбрать точку-прототип первого класса (например, объект X_1 из обучающей выборки). Количество классов K положить равным 1. Обозначить точку-прототип Z_1.
2. Определить наиболее удаленный от Z_1 объект X_f по условию

$$D(Z_1,X_f) = \max D(Z_1, X_i),$$

где $D(Z_1,X_f)$ - расстояние между Z_1 и X_f, вычисленное одним из возможных способов. Объявить X_f прототипом второго класса. Обозначить X_f как Z_2. Число классов $K = K + 1$.

Алгоритм

3. Определить пороговое расстояние T = $D(Z_1,Z_2)/2$.

Построить $\tilde{X}' = \tilde{X} \setminus \{Z_1, Z_2\}$

4. Выбрать $X_j \in \tilde{X}'$.

5. Вычислить расстояние от X_j до всех точек-прототипов: $D(Z_k, X_j)$ для $k = 1, 2, \ldots K$.

6. Определить ближайшую к рассматриваемому объекту точку-прототип Z_p по условию

$$D(Z_p, X_j) = \min_k D(Z_k, X_j).$$

Алгоритм

7. Если $D(Z_p, X_j) < T$, отнести объект X_j к классу p (Z_p является прототипом этого класса) и удалить его из $\tilde{X}'$. Иначе объявить X_j прототипом нового класса. Обозначить X_j как Z_{K+1}. Число классов K увеличить на 1: $K = K + 1$. Удалить Xj из $\tilde{X}'$.
8. Если $\tilde{X}' = \emptyset$ (то есть обучающее множество исчерпано), то КОНЕЦ. В противном случае перейти к шагу 4.

Рассмотрим пример работы алгоритма, основанного на вычислении порогового расстояния. Пусть каждый объект из множества объектов, представленных в таблице, задан двумя признаками (модель - точка на плоскости)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Координаты объектов	(2;2)	(3;3)	(2;4)	(7;1)	(8;5)	(9;1)	(9;2)	(9;6)

Выберем в качестве точки-прототипа первого класса точку X_1 из обучающей выборки (обозначается далее Z_1). В таблице представлены расстояния от этой точки до объектов X_2 — X_8.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Координаты объектов	(2;2)	(3;3)	(2;4)	(7;1)	(8;5)	(9;1)	(9;2)	(9;6)
Точки прототипы	Z_1							
Расстояние от Z_1		1,4	2	5,1	6,7	7,05	7,0	8,05

Наиболее удаленным объектом для Z_1 будет X_8.

Пороговое расстояние

$$T = \frac{1}{2} D(Z_1, X_8) = \frac{1}{2}\sqrt{(9-2)^2 + (6-2)^2} = 4,02$$

Точка X_8 становится точкой-прототипом второго класса и обозначается далее Z_2.

Рассматриваем точки множества

$\tilde{X}' = \tilde{X} \setminus \{X_1, X_f\}$. Это точки $X_2 - X_7$. Анализируем их последовательно. Точки X_2 и X_3 будут отнесены к классу 1 (прототип – Z_1).

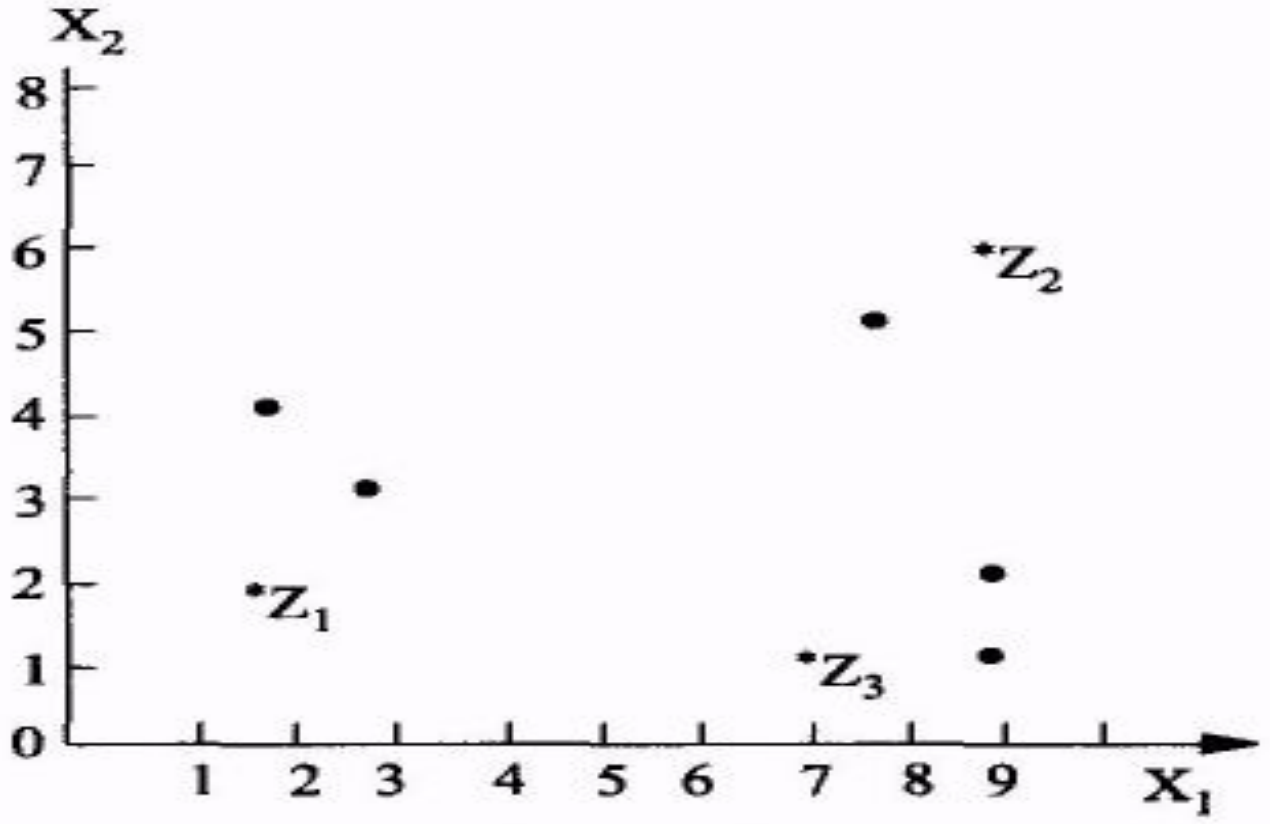
Точка X_4 имеет ближайшим прототипом Z_1, однако, поскольку расстояние $D(Z_1, X_4) > T$, точка X_4 становится прототипом нового (третьего) класса (обозначаем ее далее Z_3).

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Координаты объектов	(2;2)	(3;3)	(2;4)	(7;1)	(8;5)	(9;1)	(9;2)	(9;6)
Точки прототипы	Z_1							Z_2
Расстояние от Z_1		1,4	2	5,1	6,7	7,05	7,0	8,05
Расстояние от Z_2		6,7	7,3	5,4	1,4	5	4	

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Координаты объектов	(2;2)	(3;3)	(2;4)	(7;1)	(8;5)	(9;1)	(9;2)	(9;6)
Точки прототипы	Z_1			Z_3				Z_2
Расстояние от Z_1		1,4	2	5,1	6,7	7,05	7,0	8,05
Расстояние от Z_2		6,7	7,3	5,4	1,4	5	4	
Расстояние от Z_3					4,1	2	2,2	

Для X_5, X_6 и X_7 ближайшими прототипами станут, соответственно, Z_2, Z_3, Z_3. Поскольку условие $D(Z_p, X_j) < T$ не нарушается (наименьшие расстояния — 1,4; 2; 2,2), новых классов не возникнет.

	Прототип	Элементы класса	
Класс 1	X_1	X_2	X_3
	(2; 2)	(3; 3)	(2; 4)
Класс 2	X_8	X_5	
	(9; 6)	(8; 5)	
Класс 3	X_4	X_6	X_7
	(7; 1)	(9; 1)	(9; 2)


X2
8
7
6
5
4
3
2
1
0
Z2
Z1
Z3
1
2
3
4
5
6
7
8
9
X1

К достоинствам рассмотренного алгоритма следует отнести простоту реализации и небольшой объем вычислений.

Недостатки:

- не предусмотрено уточнение разбиения. В результате расстояние от объекта до точки-прототипа класса может оказаться больше, чем расстояние от этого объекта до точки-прототипа другого класса.
- Результат, кроме того, сильно зависит от порядка рассмотрения объектов *X,* а также от способа вычисления порогового расстояния (можно использовать и другие формулы для подсчета *T).*

Из этого следует, что полезно было бы использовать алгоритмы, допускающие многократную коррекцию формируемых классов, например, можно было бы менять пороговое расстояние T и проводить многократное уточнение разбиения.

Алгоритм MAXMIN

Рассмотрим алгоритм, более эффективный по сравнению с предыдущим и являющийся улучшением порогового алгоритма. Исходными даннымы для работы алгоритма будет, как и раньше, выборка X. Объекты этой выборки следует разделить на классы, число и характеристики которых заранее неизвестны.

Алгоритм MAXMIN

На первом этапе алгоритма все объекты разделяются по классам на основе критерия минимального расстояния от точек-прототипов этих классов (первая точка-прототип может выбираться произвольо). Затем в каждом классе выбирается объект, наиболее удаленный от своего прототипа. Если он удален от своего прототипа на расстояние, превышающее пороговое, такой объект становится прототипом нового класса.

В этом алгоритме пороговое расстояние не является фиксированным, а определяется на основе среднего расстояния между всеми точками-прототипами, то есть корректируется в процессе работы алгоритма. Если в ходе распределения объектов выборки *X* по классам были созданы новые прототипы, процесс распределения повторяется. Таким образом, в алгоритме MAXMIN окончательным считается разбиение, для которого в каждом классе расстояние от точки-прототипа до всех объектов этого класса не превышает финального значения порога *T.*

Алгоритм

1. Выбрать точку-прототип первого класса (например, объект X_1 из обучающей выборки). Количество классов K положить равным 1. Обозначить точку-прототип Z_1.
2. Определить наиболее удаленный от Z_1 объект X_f по условию

$$D(Z_1,X_f) = max\ D(Z_1, X_i),$$

где $D(Z_1,X_f)$ - расстояние между Z_1 и X_f, вычисленное одним из возможных способов. Объявить X_f прототипом второго класса. Обозначить X_f как Z_2. Число классов $K = K + 1$.

Алгоритм

3. Находим пороговое расстояние T.
4. Для всех объектов обучающего множества строится матрица расстояний до каждого из имеющихся прототипов: $D(X_i, Z_k),\ i = 1, \ldots M,\ k = 1, \ldots, K$.
5. Каждый объект относится к классу по критерию наибольшей близости к точке-прототипу: X_i *отнесен к классу* p, *если* $D(X_i, Z_p) = \min\limits_k D(X_i, Z_k)$.
6. В каждом классе k определяется объект X_{lk}, наиболее удаленный от точки-прототипа: $D(X_{lk}, Z_k) = \max\limits_{X \in R_k} D(X, Z_k)$, где R_k — множество объектов класса k.

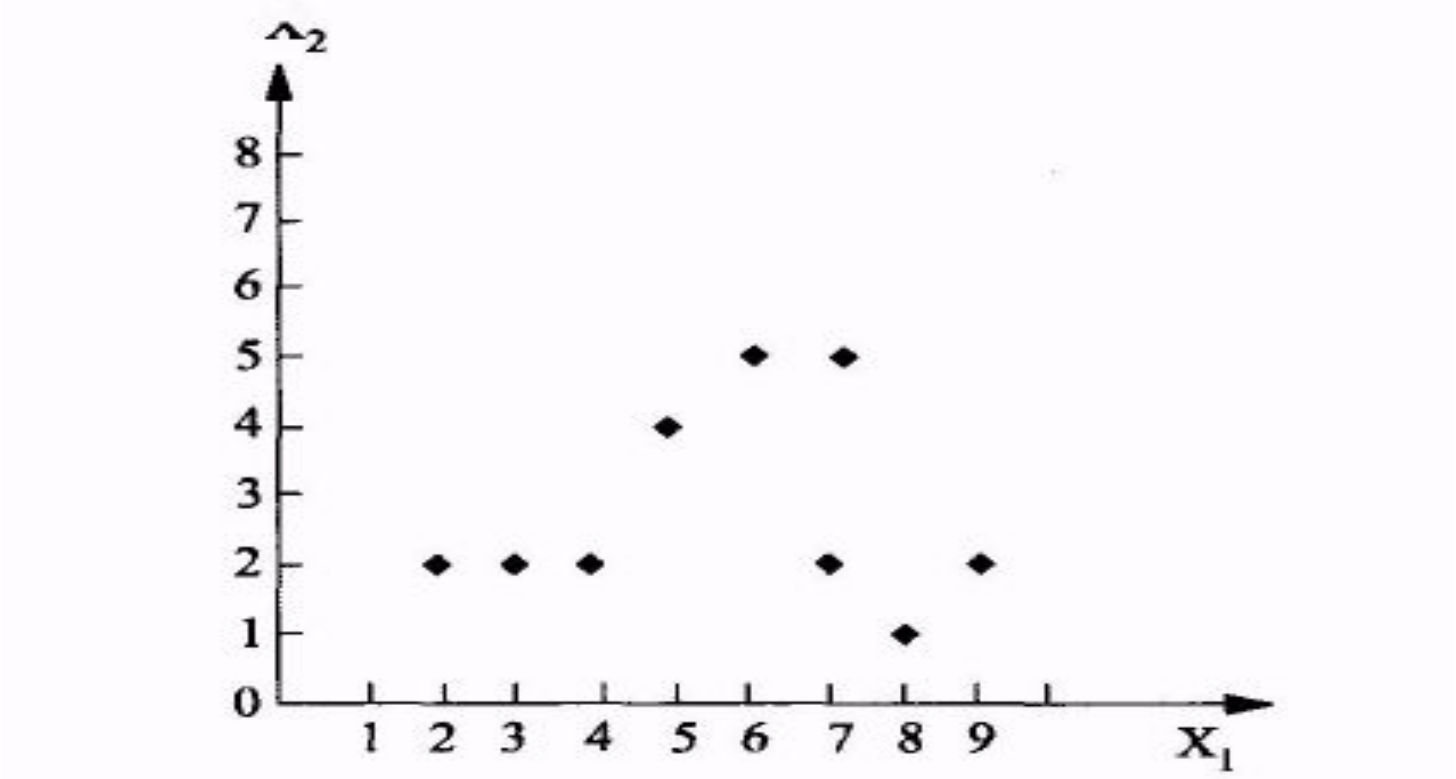
Алгоритм

7. Для всех найденных объектов проверяется условие: $D(X_{lk}, Z_k) < T, k = 1, \ldots, K$.

   Если для некоторого X_{lk} это условие не выполнено, он становится точкой-прототипом нового класса, число классов $K = K + 1$.

8. Если новых классов не создано, то КОНЕЦ. Иначе — перейти к шагу 9.
9. Вычисляется новое значение T как среднее расстояние между прототипами.
10. Перейти к шагу 4.

Рассмотрим работу алгоритма MAXMIN на примере. Как и в предыдущем случае выберем объекты, которые заданы двумя признаками. Обучающая выборка представлена на рис.



	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Координаты точек	(2; 2)	(3; 2)	(4; 2)	(5; 4)	(6; 5)	(7; 2)	(7; 5)	(8; 1)	(9; 2)
Первая точка-прототип	$\boldsymbol{Z}_1$								
Расстояние $D(Z_1, X_j)$		1	2	3,6	5	5	5,8	6,1	7