Кластерный анализ

- Классификация является одним из фундаментальных процессов в науке. . необходимый нам для понимания природы.
- Классификация – это упорядочение объектов по схожести.
- Понятие схожести является неоднозначным

- Группы близких по какому-либо критерию объектов обычно называются кластерами.
- Необходимость в использовании методов кластерного анализа возникает в том случае, когда задано множество характеристик, по которым тестируется множество испытуемых; задача состоит в выделении классов (групп) испытуемых, близких по всему множеству характеристик (профилю).

Определение

Кластерный анализ (*англ. cluster analysis)* - математическая процедура многомерного анализа, позволяющая н***а основе множества показателей*** (как объективных, так и субъективных), характеризующих ряд ***объектов*** (напр., испытуемые, стимулы), ***сгруппировать их в классы*** (кластеры) т.о., чтобы объекты, входящие в один класс, были более однородными, сходными по сравнению с объектами, входящими в др. классы.

Факторный и кластерный анализ

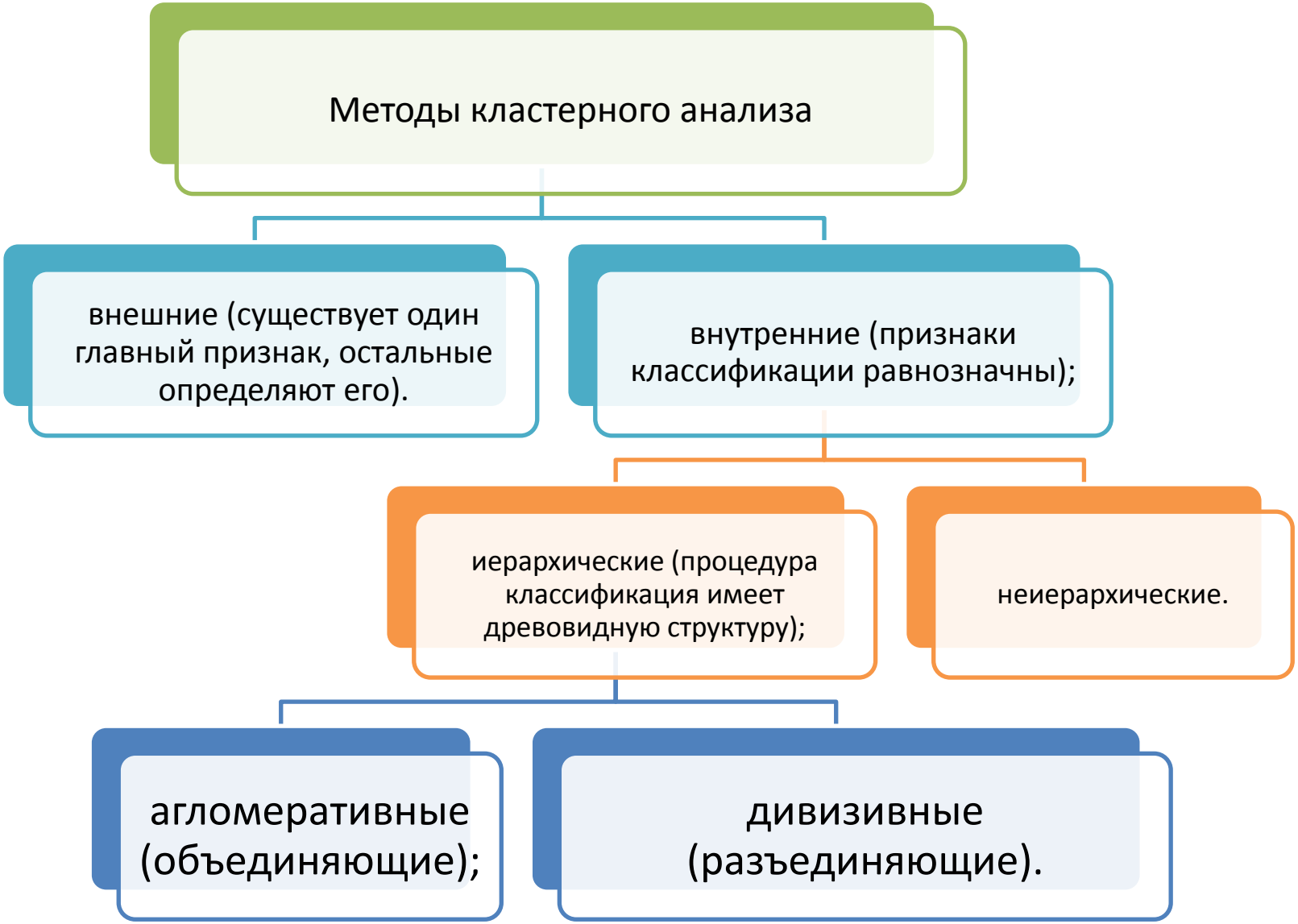
- Кластерный и факторный анализы преследуют цель: классифицировать переменные и/или категории респондентов по однородным группам (сегментам, кластерам).
- Принципы классификации также могут быть различными. Поэтому часто процедуры, используемые в кластерном анализе для формирования классов, основываются на фундаментальных процессах классификации, присущих людям и другим живым существам

Пример

Темперамент

- Показатели
  - Факторный анализ
- Группы – типы темперамента
  - Кластерный анализ

- Необходимость в использовании методов кластерного анализа возникает в том случае, когда **задано множество характеристик**, по которым тестируется **множество испытуемых;**
- задача состоит **в выделении классов (групп) испытуемых**, близких по всему множеству характеристик (профилю).
- На первом этапе матрица смешения (оценки людей по различным характеристикам) преобразуется в матрицу расстояний.
- Для подсчета матрицы расстояния осуществляется подбор метрики, или метода вычисления расстояния между объектами в многомерном пространстве. Если объект описывается *k* признаками, то он может быть представлен как точка в *k*-мерном пространстве. Возможность измерения расстояний между объектами в *k*-мерном пространстве вводится через понятие метрики.


Методы кластерного анализа
внешние (существует один главный признак, остальные определяют его).
внутренние (признаки классификации равнозначны);
иерархические (процедура классификация имеет древовидную структуру);
неиерархические.
агломеративные (объединяющие);
дивизивные (разъединяющие).

- В *иерархических* методах выстраивается «дерево» кластеров, то есть для полученных окончательных кластеров можно проследить «историю» их постепенного формирования путем объединения или разъединения первоначально существовавших кластеров

- В *итеративных* методах разбиение на кластеры получается из некоторого начального разбиения способом последовательных перерасчетов (приближений, итераций).

Дивизивные методы

- В *дивизивных* иерархических методах множество исходных данных первоначально представляется как один кластер, который затем разделяется на некоторое (часто заранее заданное) количество кластеров.
- Процесс кластеризации заканчивается, когда получено разделение исходного множества данных на заданное число кластеров при определенном удовлетворяющем исследователя качестве разделения.

Агломеративные методы

- В *агломеративных* иерархических методах, каждый элемент (результат измерения) эмпирической выборки первоначально представляется отдельным кластером.
- Затем эти кластеры начинают объединять; при этом на каждом шаге кластеризации объединяются наиболее близкие друг к другу кластеры. Новые полученные образования представляют собой кластеры более высокого уровня в иерархии кластеров, именно поэтому такие методы часто называют методами иерархической кластеризации.
- Процесс кластеризации обязательно заканчивается за конечное число шагов, так как в итоге все данные оказываются объединенными в один-единственный кластер, совпадающий со всей исходной эмпирической выборкой.

- Варианты кластерного анализа — это **множество простых вычислительных процедур**, используемых для классификации объектов.
- *Классификация* объектов — это группирование их в классы так, чтобы объекты в каждом классе были более похожи друг на друга, чем на объекты из других классов.
- Более точно, кластерный анализ — это процедура упорядочивания объектов в сравнительно **однородные классы на основе попарного сравнения этих объектов по предварительно определенным и измеренным критериям**

Задачи, при решении которых кластерный анализ является более эффективным

- **разбиение совокупности испытуемых на группы** по измеренным признакам с целью дальнейшей **проверки причин межгрупповых различий** по внешним критериям, например, проверка гипотез о том, проявляются ли **типологические различия** между испытуемыми по измеренным признакам;

- применение кластерного анализа как значительно более простого и наглядного **аналога факторного анализа**, когда ставится только задача группировки признаков на основе их корреляции;

- классификация объектов на основе непосредственных оценок различий между ними (например, исследование социальной структуры коллектива по данным социометрии — по выявленным межличностным предпочтениям).

Этапы Кластерного анализа

1. Отбор объектов для кластеризации. Объектами могут быть, в зависимости от цели исследования:

а) испытуемые;

б) объекты, которые оцениваются испытуемыми;

в) признаки, измеренные на выборке испытуемых.

Этапы Кластерного анализа

2. *Определение множества переменных,* по которым будут различаться объекты кластеризации.

Для испытуемых — это набор измеренных признаков,

для оцениваемых объектов — субъекты оценки,

для признаков — испытуемые.

Если в качестве исходных данных предполагается использовать результаты попарного сравнения объектов, необходимо четко определить критерии этого сравнения испытуемыми (экспертами).

Этапы Кластерного анализа

3. Определение меры различия между объектами кластеризации.

Это первая проблема, которая является специфичной для методов анализа различий: многомерного шкалирования и кластерного анализа.

Этапы Кластерного анализа

4. Выбор и применение метода классификации для создания групп сходных объектов.

Это вторая и центральная проблема кластерного анализа.

разные методы кластеризации порождают разные группировки для одних и тех же данных.

Анализ заключается в обнаружении структуры, на деле в процессе кластеризации структура привносится в данные, и эта привнесенная структура может не совпадать с реальной.

Этапы Кластерного анализа

5. *Проверка достоверности разбиения* на классы.

(не всегда необходим,)

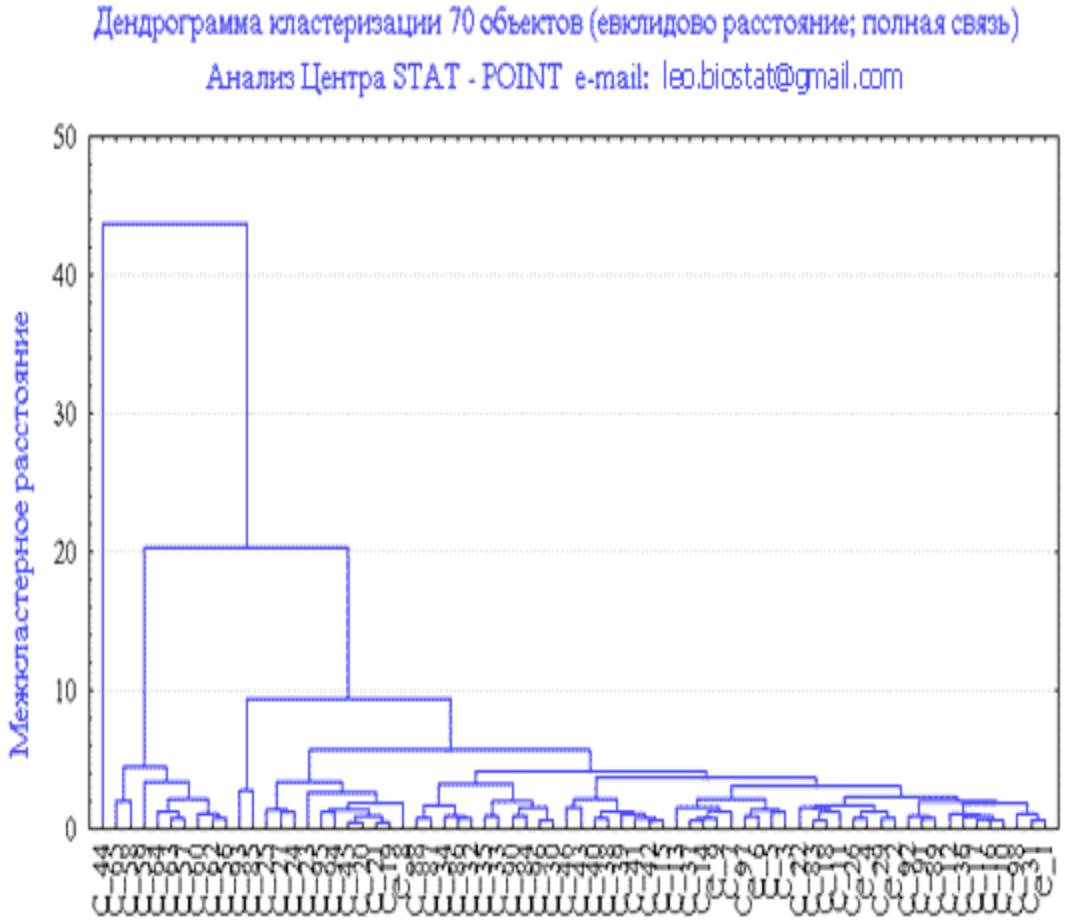
- кластерный анализ *всегда* разобьет совокупность объектов на классы, независимо оттого, существуют ли они на самом деле.
- Обычно проверяют *устойчивость группировки* — на повторной идентичной выборке объектов.
- *Значимость разбиения* проверяют по внешним критериям — признакам, не вошедшим в анализ.

Методы кластерного анализа

Критерий объединения многократно применяется ко всей матрице попарных расстояний между объектами.

- На первых шагах объединяются наиболее близкие объекты, находящиеся на одном уровне сходства. Затем поочередно присоединяются остальные объекты, пока все они не объединятся в один большой кластер.
- Результат работы метода представляется графически в виде дендрограммы — ветвистого древовидного графика.

Методы кластерного анализа



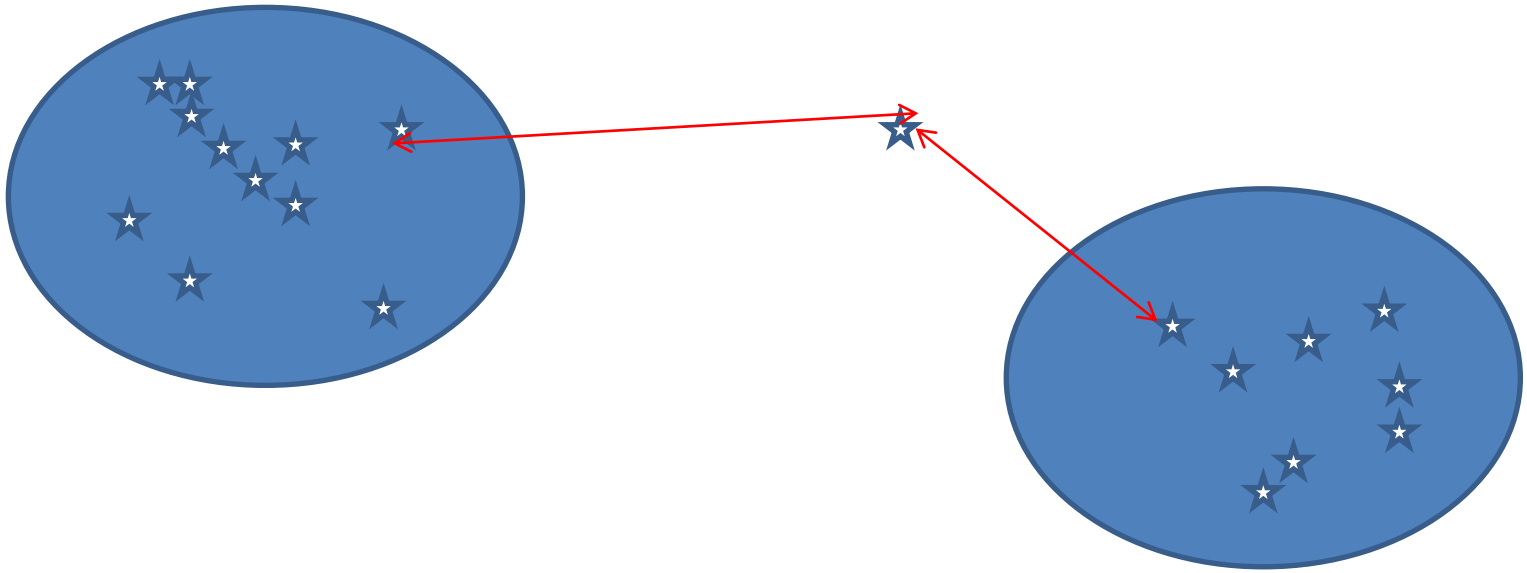
Методы кластерного анализа

- расстояние рассматривается в двух смыслах:
- 1) как расстояние между объектами внутри кластера
- 2) как расстояние между различными кластерами, получаемыми в процессе кластеризации, или, другими словами, как ***межкластерное расстояние***.

Методы кластерного анализа

- *Single linkage*, *nearest neighbor* (Простая связь, или метод «ближнего соседа») – расстояние между двумя кластерами определяется как попарное расстояние между двумя самыми ближними друг к другу представителям каждого из них. Метод простой связи сильно сжимает исходное признаковое пространство и рекомендуется для получения минимального «дерева» объединения

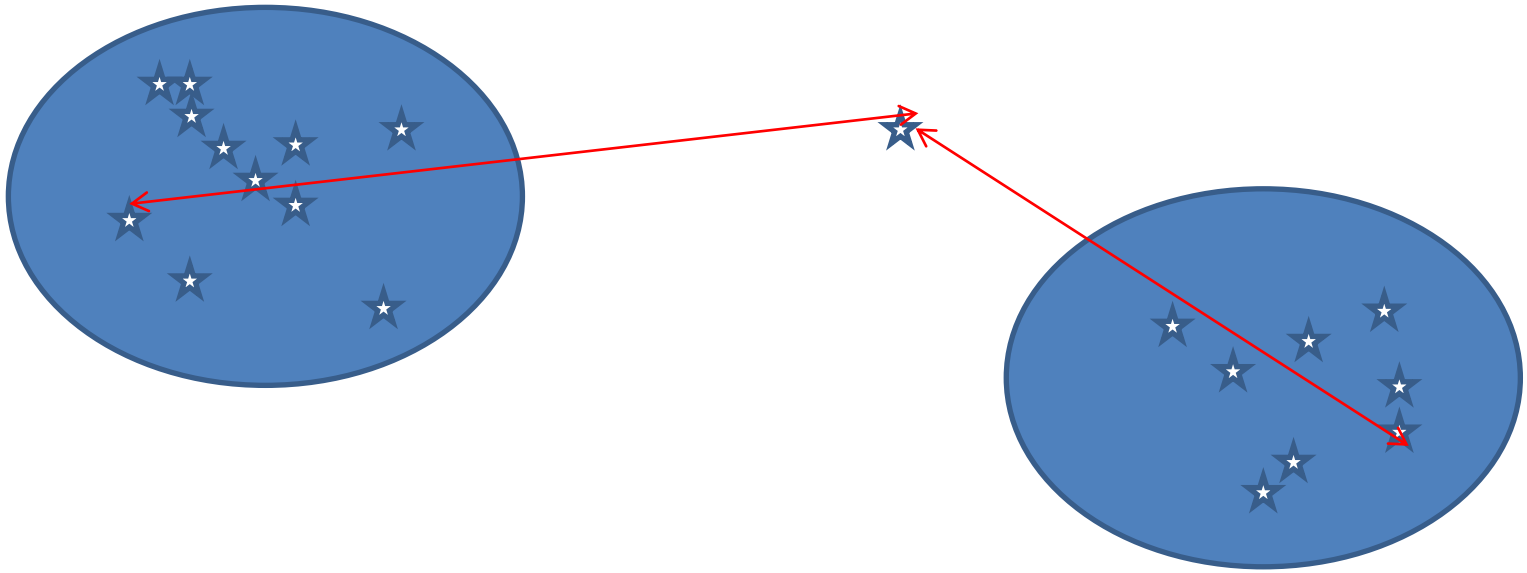
Методы кластерного анализа



Методы кластерного анализа

- *Complete linkage*, *furthest neighbor* (Полная связь, или метод «дальнего соседа») – расстояние между двумя кластерами определяется по самым дальним друг от друга представителям каждого из них. Этот метод сильно растягивает исходное пространство

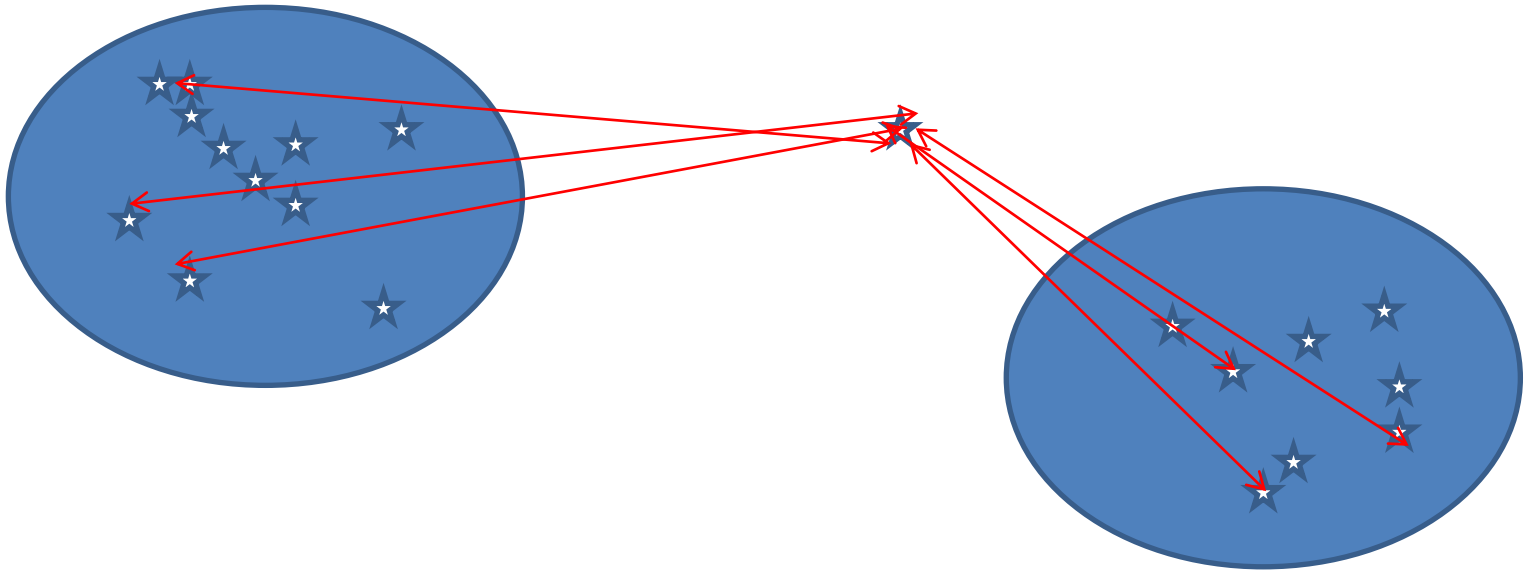
Методы кластерного анализа



Методы кластерного анализа

- Метод средней связи *(Average Linkage)* пли *межгрупповой связи (Between Groups Linkage)* занимает промежуточное положение относительно крайно-стей методой одиночной и полной связи. На каждом шаге вычисляется среднее арифметическое расстояние между каждым объектом из одного кластера и каждым объектом из другого кластера. Объект присоединяется к данному кластеру, если это среднее расстояние меньше, чем среднее расстояние до любого другого кластера. По своему принципу этот метод должен давать *более точные результаты классификации,* чем остальные методы.

Методы кластерного анализа



Кластерный анализ на основе теории Л.С.Выготского.

В работе «Мышление и речь» Л.С. Выготский описывает различные генетические ступени развития понятий

- ассоциативный комплекс,
- комплекс-коллекция,
- цепной комплекс,
- диффузный комплекс,
- псевдопонятия.

Пример

- Предмет изучения: отношения между членами некоей малой группы (производственной, научной или учебной)
- Для одной и той же группы может быть выделено несколько типов отношений:
  - производственные,
  - личные,
  - общность увлечений и т.д.

- Тогда для какой-либо из групп экспериментально определяется структура отношений каждого типа и строится матрица попарных расстояний (или близости) между членами группы по каждому типу отношений.

Количество членов малой группы, т.е. элементов рассматриваемого множества, n=9

m=3 различных типов отношений между членами малой группы:

1) взаимоотношения, связанные с основной работой,

2) взаимоотношения, связанные с неделовыми формами общения,

3) взаимоотношения, связанные с участием в дополнительной работе.

Алгоритмы образования комплексов различных типов

1.Ассоциативный *кластер.*

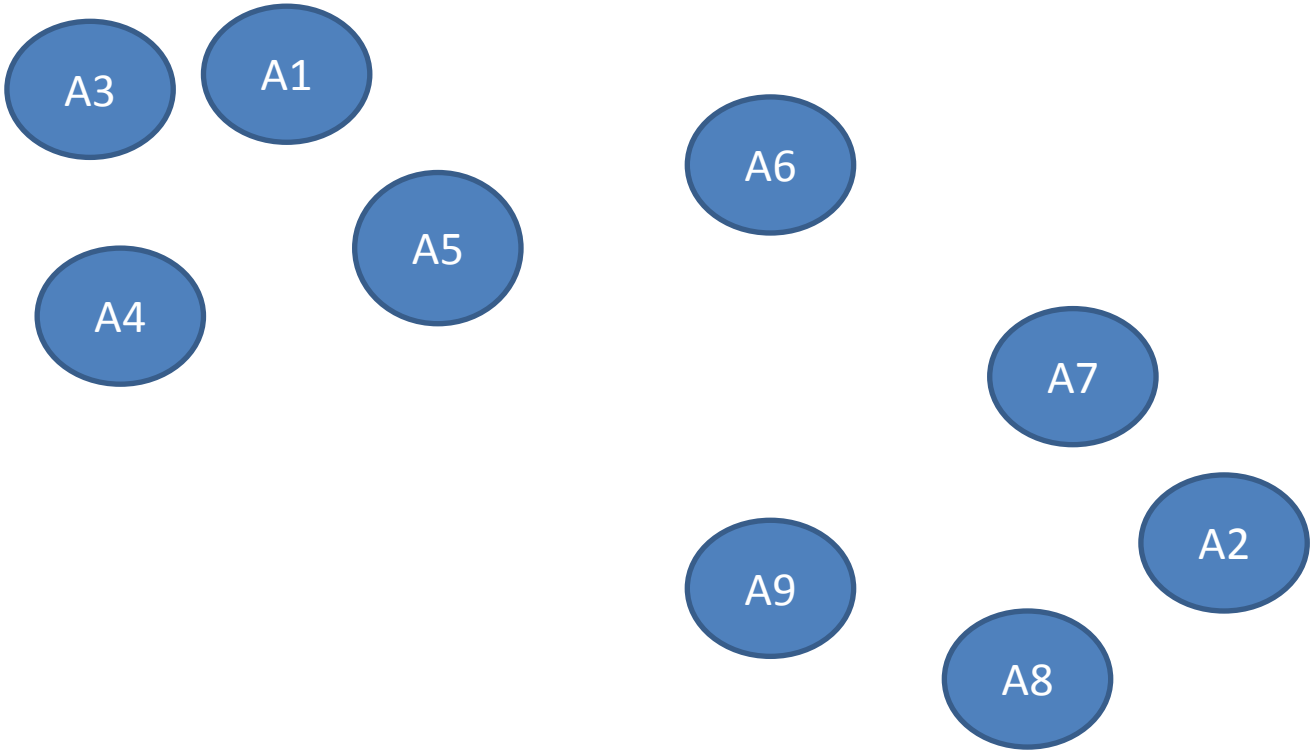
Сначала выбирается множество элементов, которые в совокупности будут составлять ядро обобщенного ассоциативного кластера.

Далее по каждому признаку для каждого из элементов ядра отбираются ближайшие по выбранному признаку элементы, а величины этих минимальных расстояний фиксируются.

Затем из всех расстояний выбирается наименьшее, и происходит отбор только тех элементов, которые находятся на минимальном расстоянии от какого-либо из элементов ядра. Эта процедура повторяется для всех качеств.

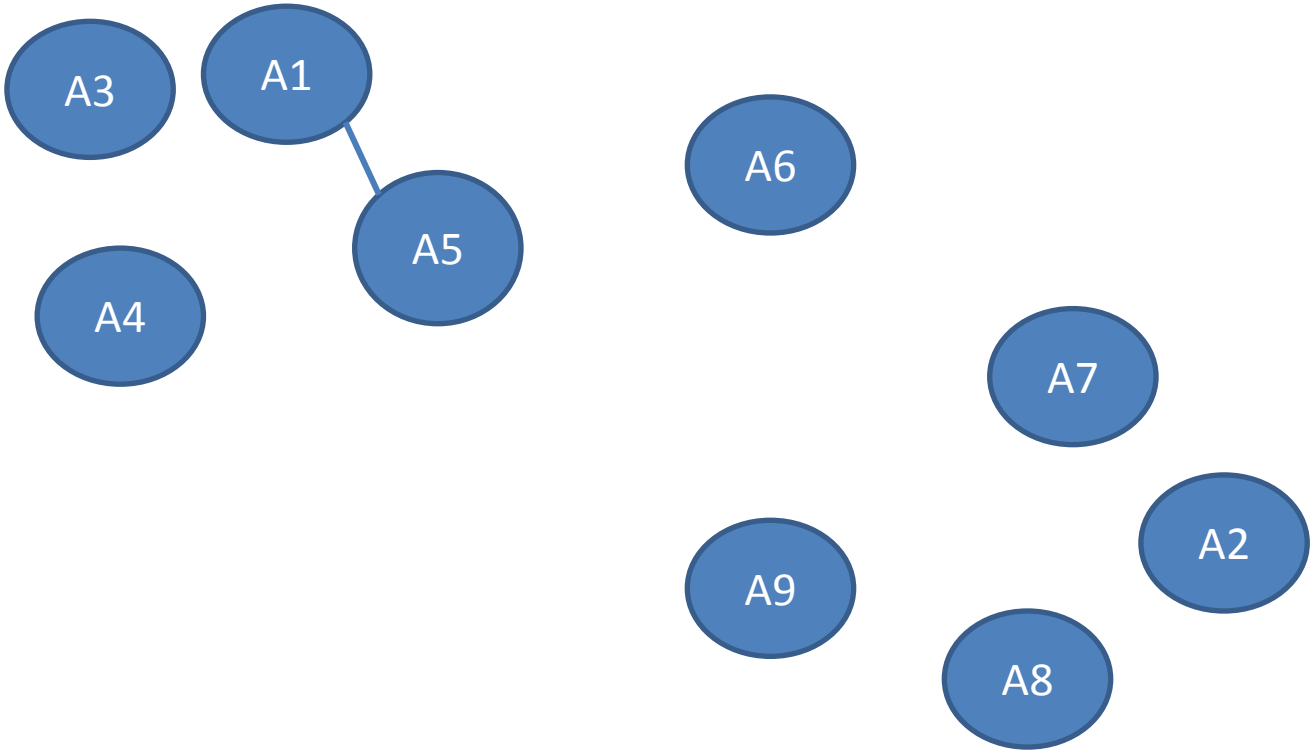
Алгоритмы образования комплексов различных типов

1.Ассоциативный *кластер.*



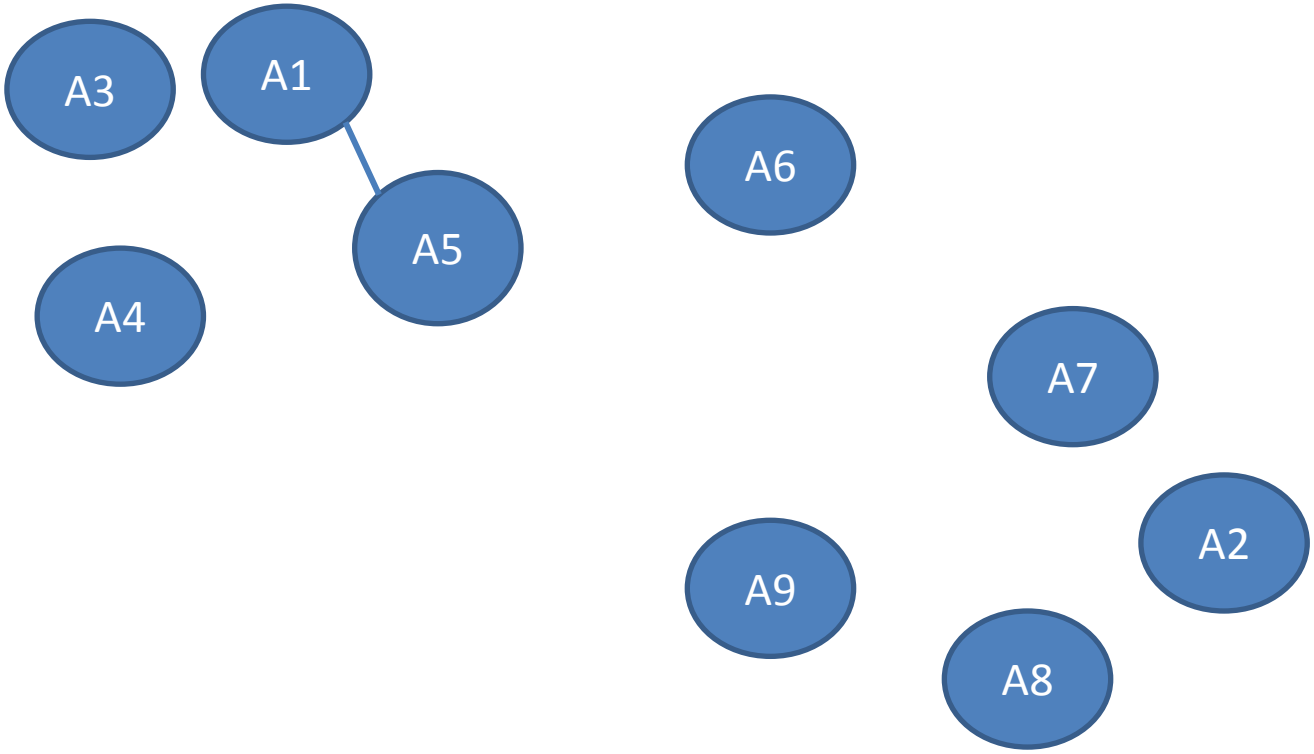
Алгоритмы образования комплексов различных типов

1.Ассоциативный *кластер.*



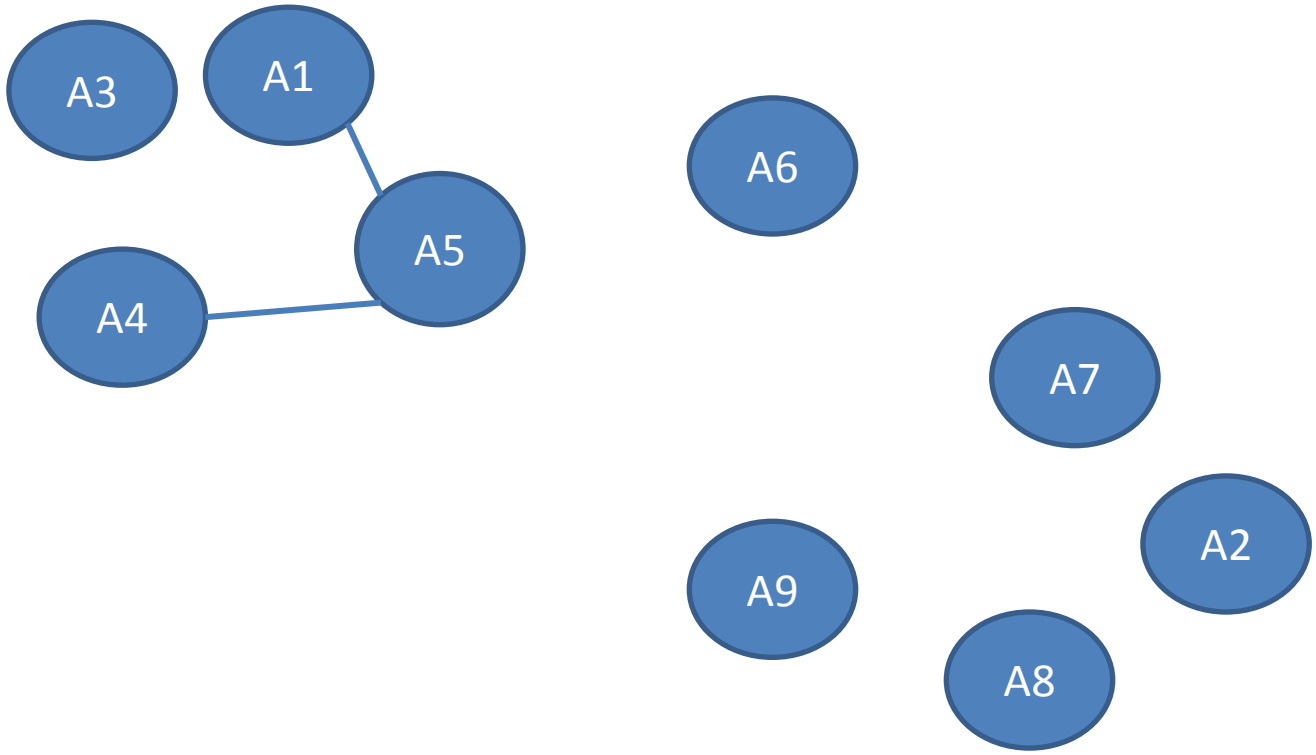
Алгоритмы образования комплексов различных типов

1.Ассоциативный *кластер.*



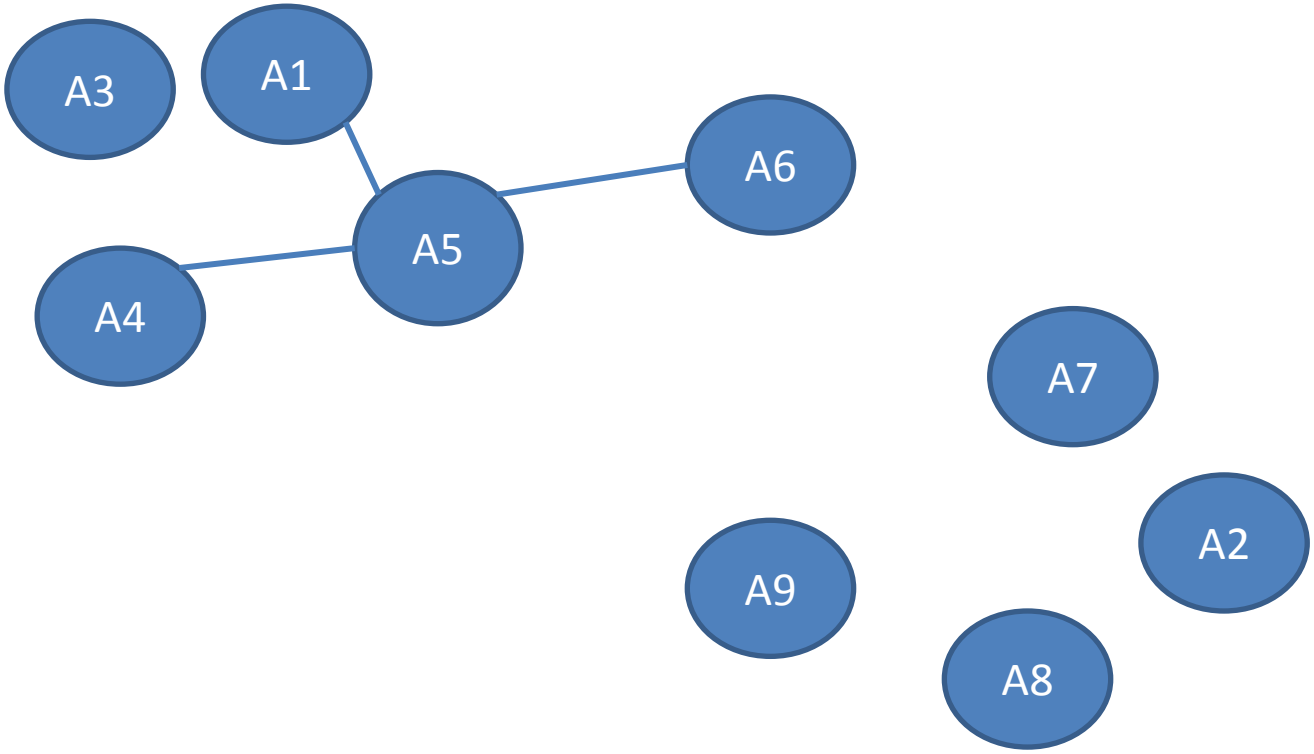
Алгоритмы образования комплексов различных типов

1.Ассоциативный *кластер.*



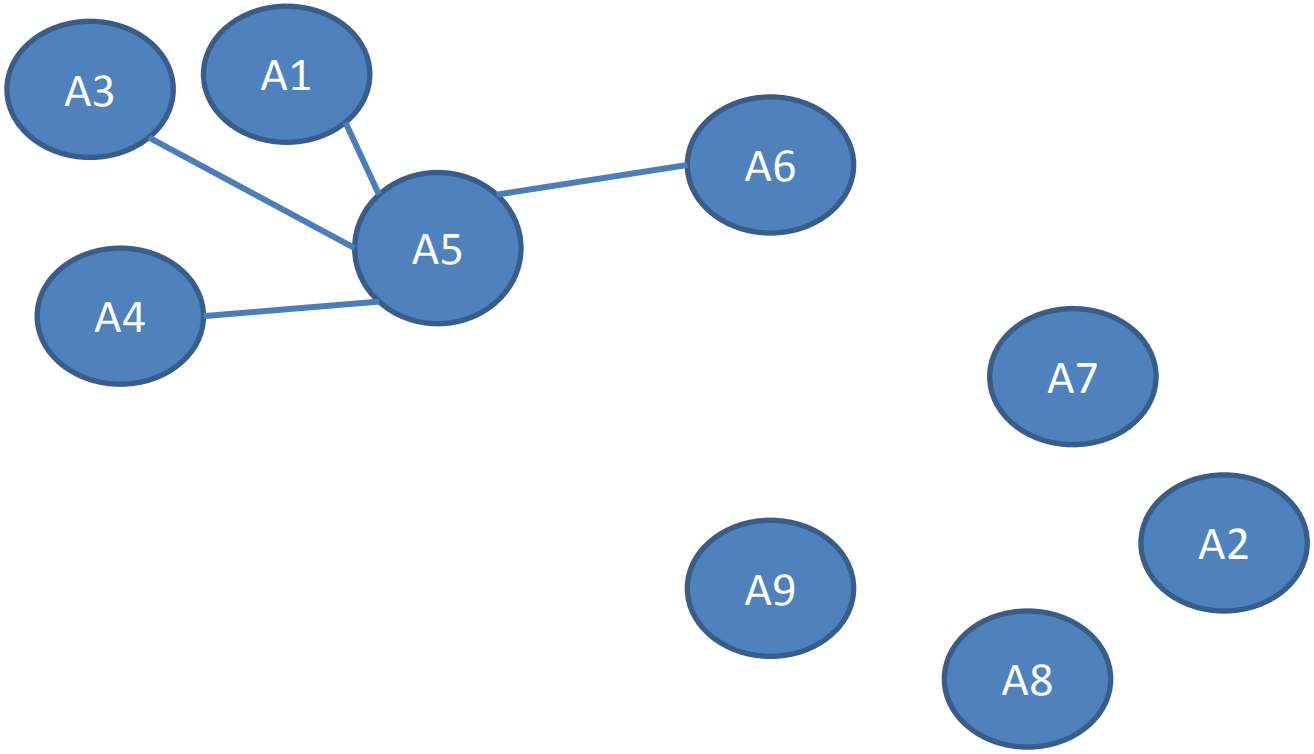
Алгоритмы образования комплексов различных типов

1.Ассоциативный *кластер.*



Алгоритмы образования комплексов различных типов

1.Ассоциативный *кластер.*



1.Ассоциативный *кластер.*

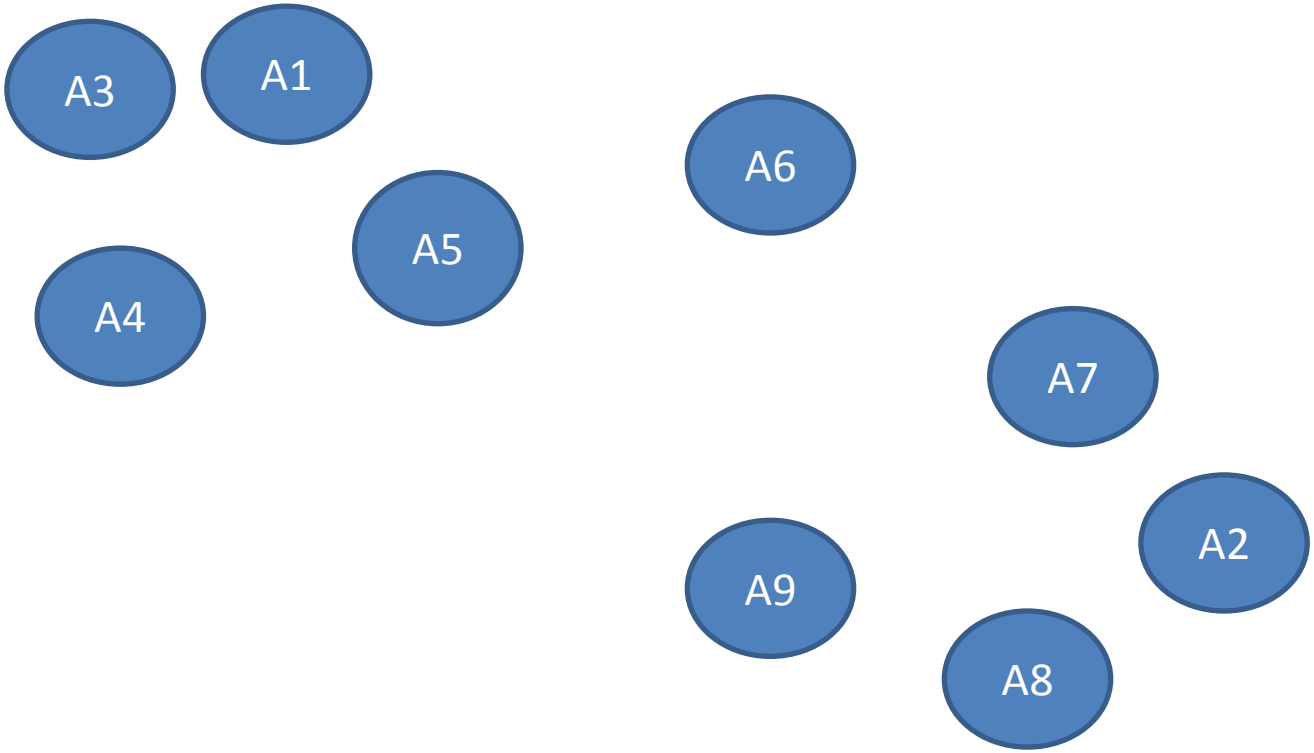
- . Элементы ассоциативного комплекса (по Выготскому) могут вовсе не быть объединены между собой, а находиться в ассоциативной связи лишь с ядром комплекса. Это означает, что apriori могут быть заданы не все расстояния, т.е. множество элементов упорядочится лишь частично.

2. *Цепной кластер*

«***Цепной комплекс строится по принципу динамического временного объединения отдельных звеньев в единую цепь и переноса значения через отдельные звенья этой цепи. Каждое звено соединено... с предшествующим... (и)... последующим, причем самое важное отличие этого типа комплекса в том, что характер связи или способ соединения одного и того же звена с предшествующим и последующим может быть совершенно различным***»

Алгоритмы образования комплексов различных типов

2.Цепной *кластер.*

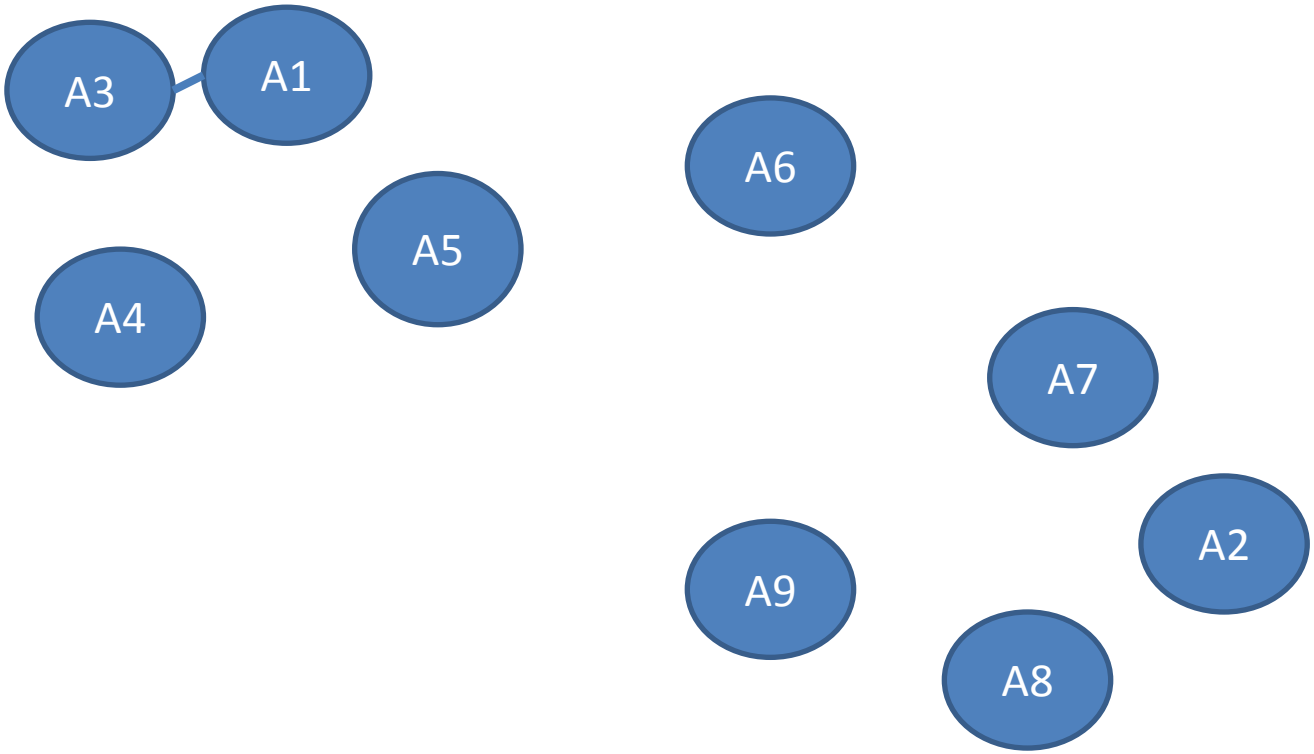


Сначала из заданного множества n элементов выбирается один, который станет первым элементом, составляющим цепной кластер.

Затем для каждого качества (т.е. для каждой матрицы расстояний из *m* заданных матриц) выбирается элемент, ближайший к первому. Из полученных *M* минимальных расстояний выбирается наименьшее и фиксируется номер соответствующей матрицы и номер элемента – этот элемент и будет вторым в цепном кластере.

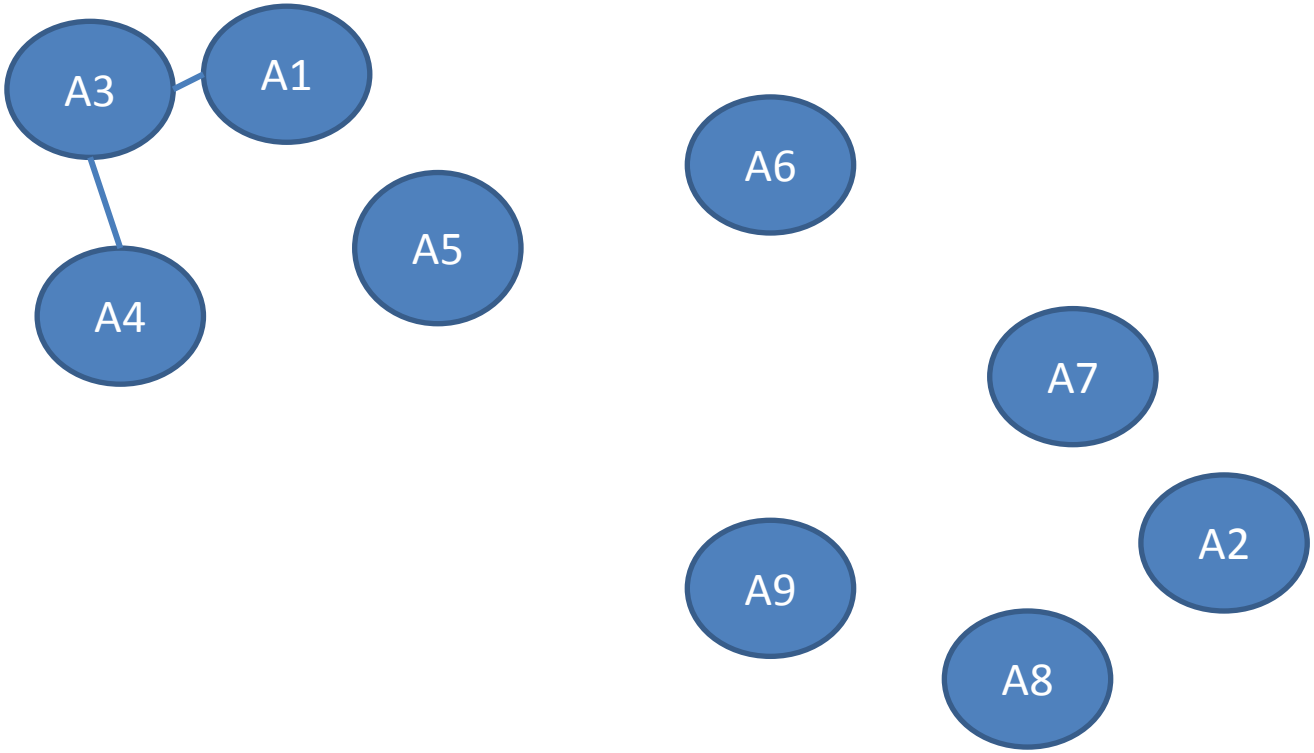
Алгоритмы образования комплексов различных типов

2.Цепной *кластер.*



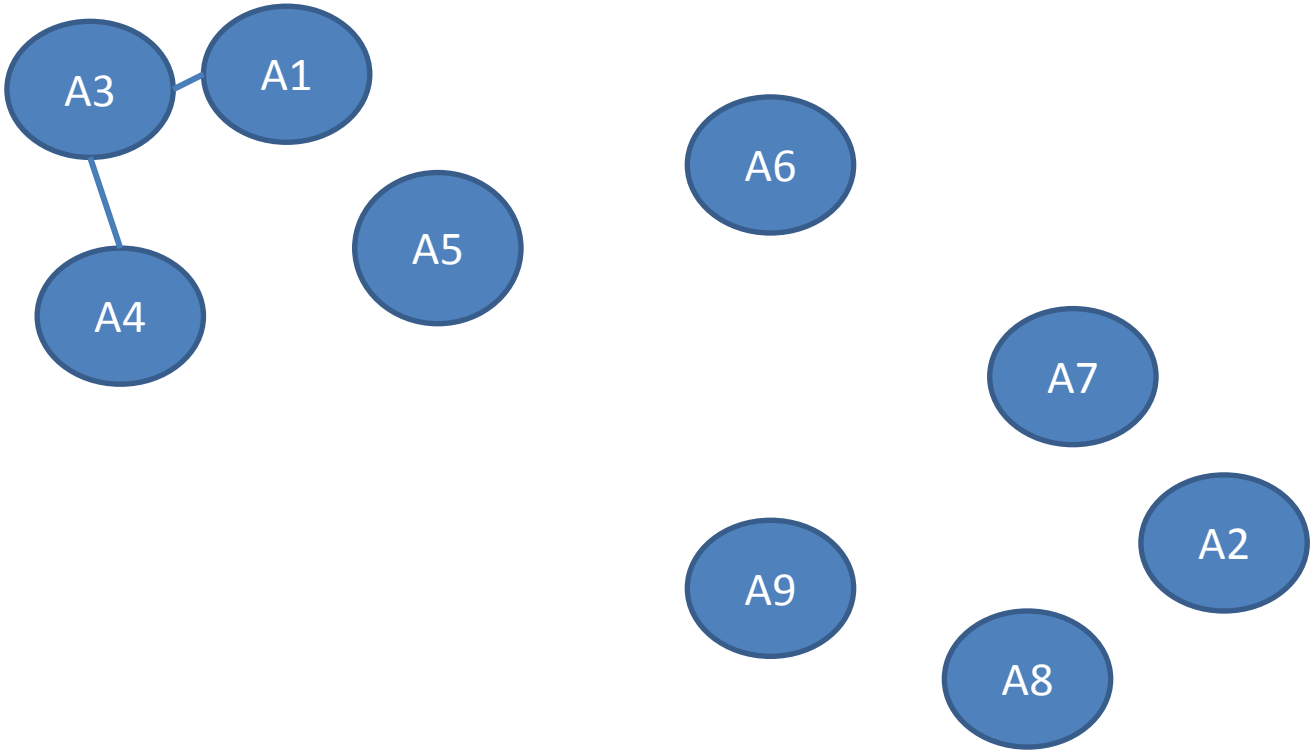
Алгоритмы образования комплексов различных типов

2.Цепной *кластер.*



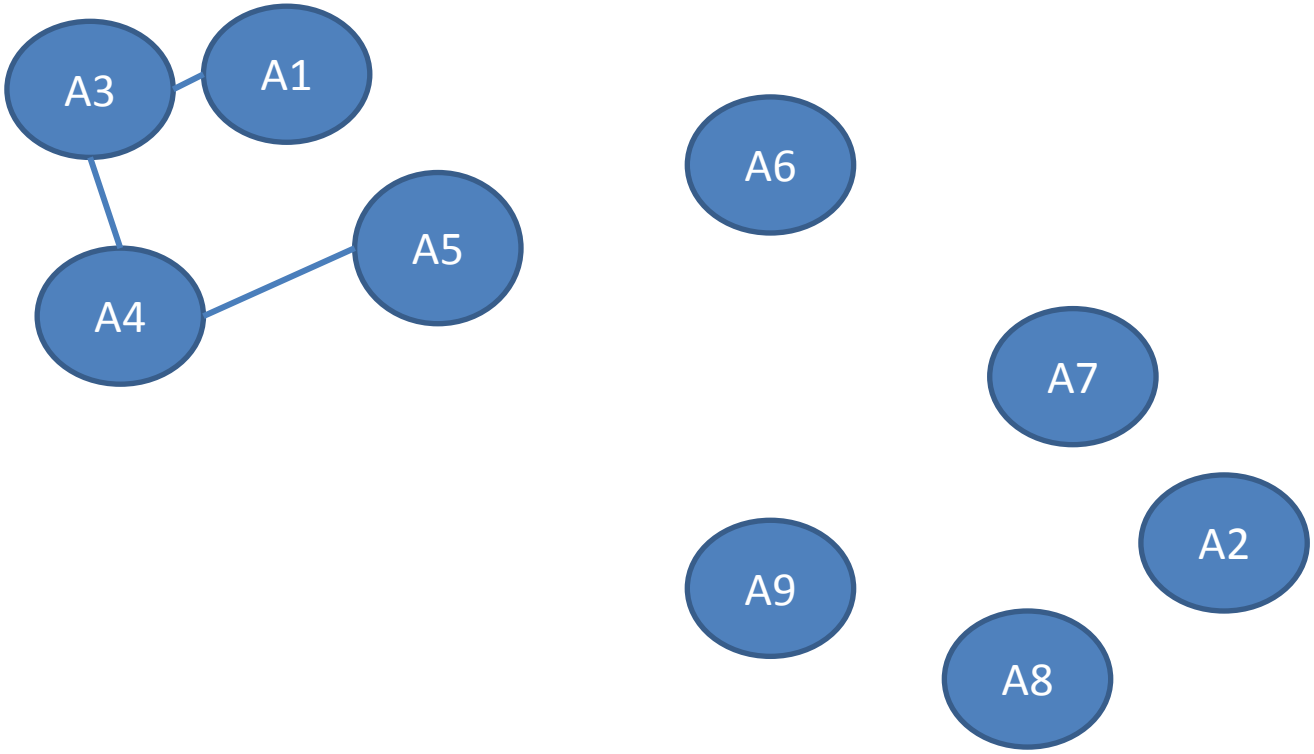
Алгоритмы образования комплексов различных типов

2.Цепной *кластер.*



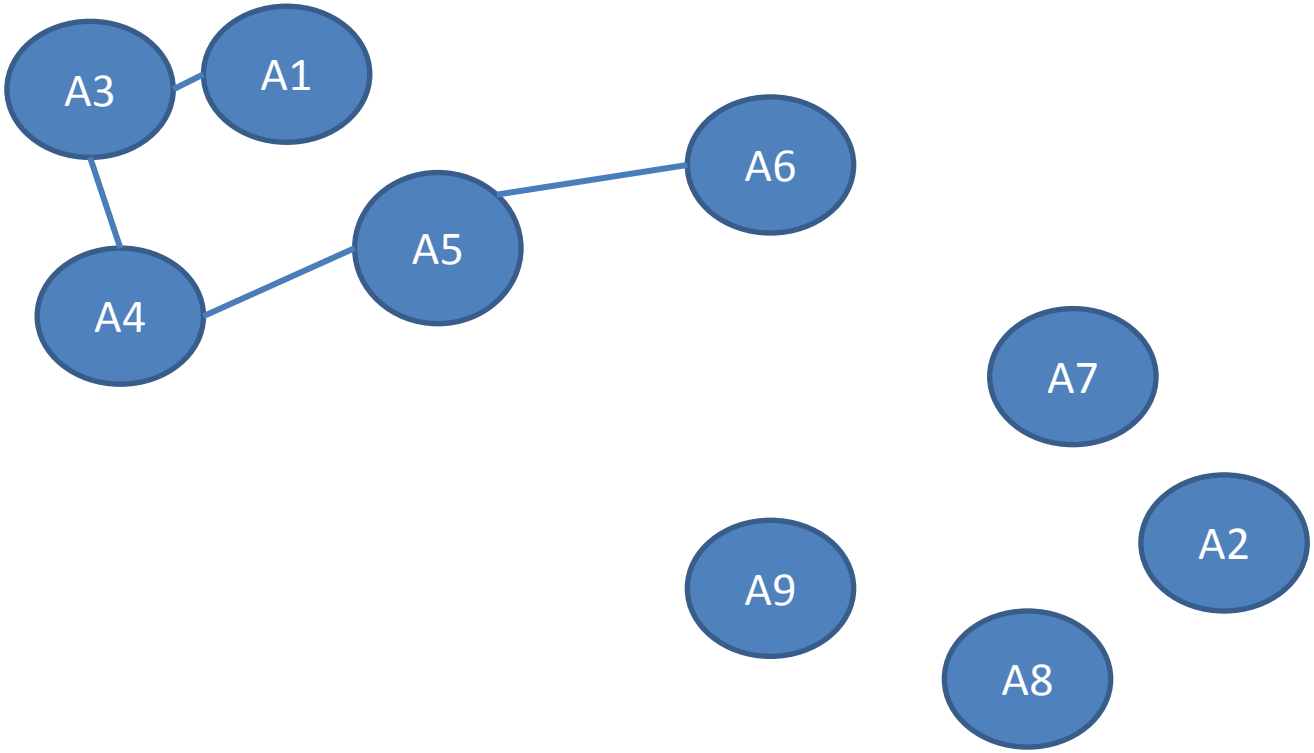
Алгоритмы образования комплексов различных типов

2.Цепной *кластер.*



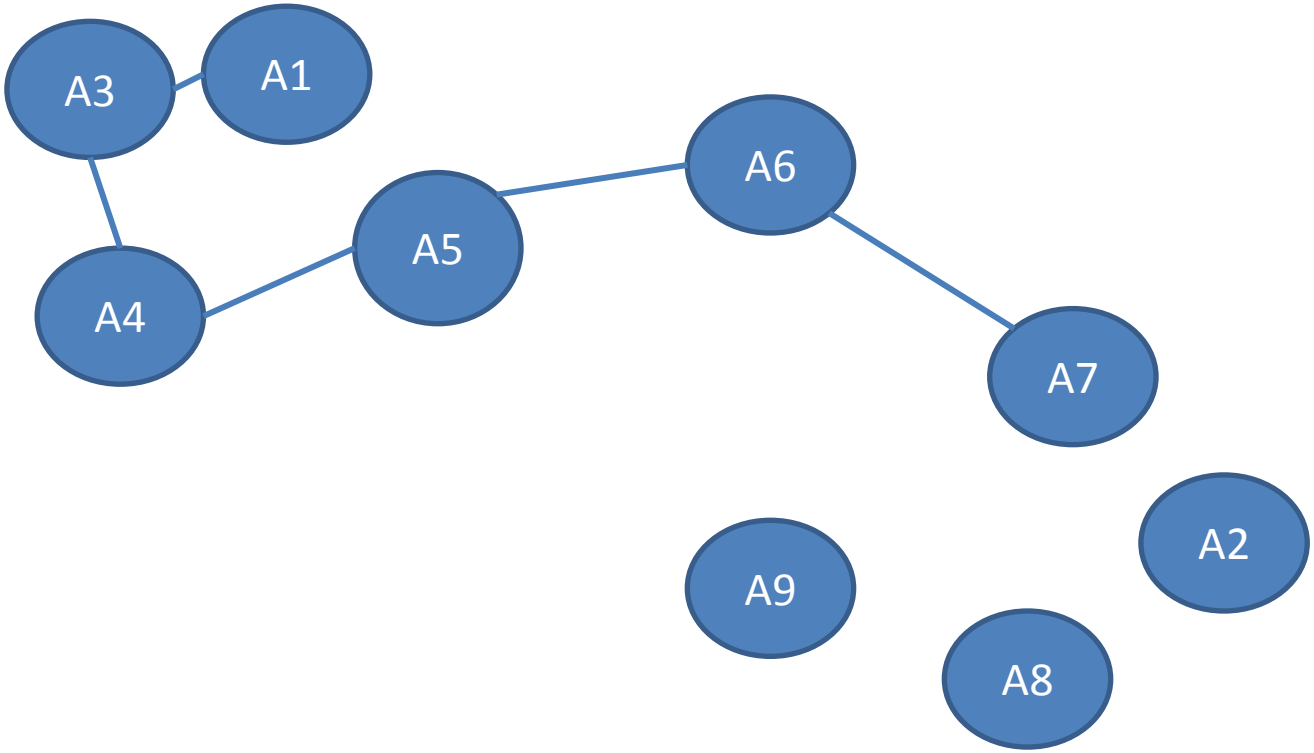
Алгоритмы образования комплексов различных типов

2.Цепной *кластер.*



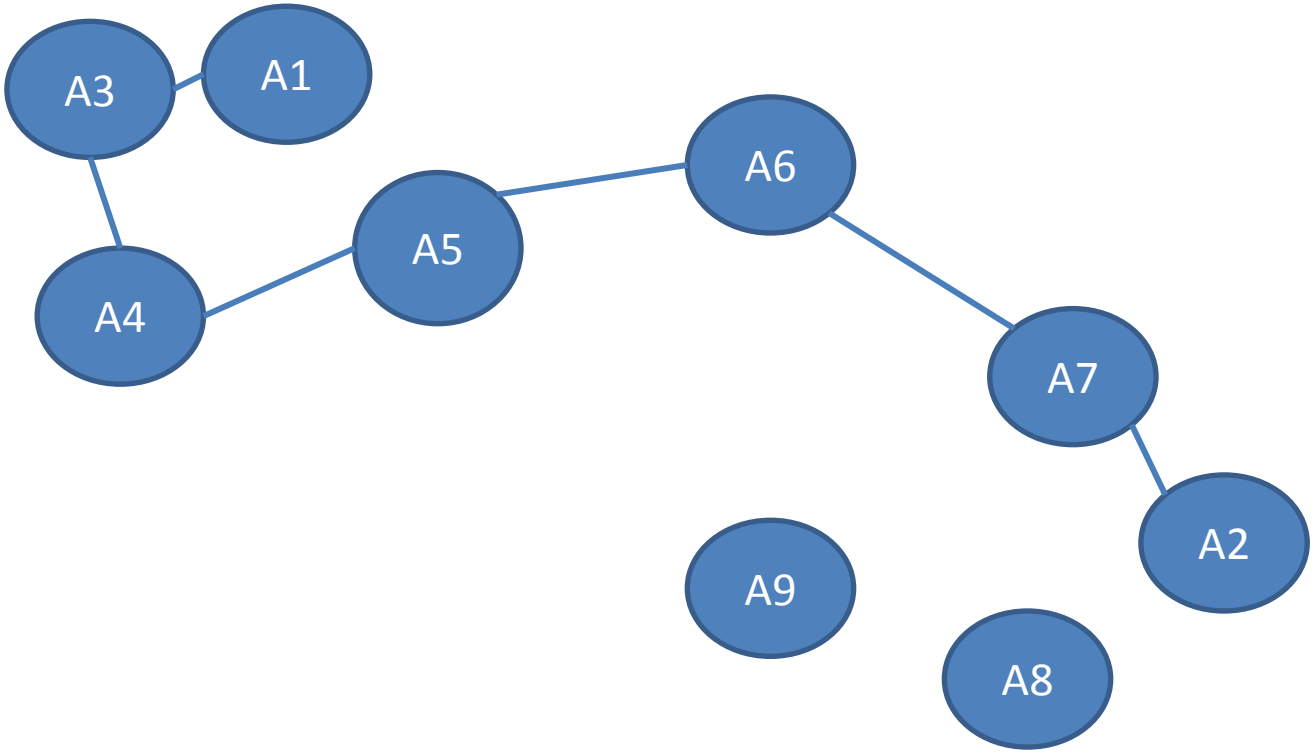
Алгоритмы образования комплексов различных типов

2.Цепной *кластер.*



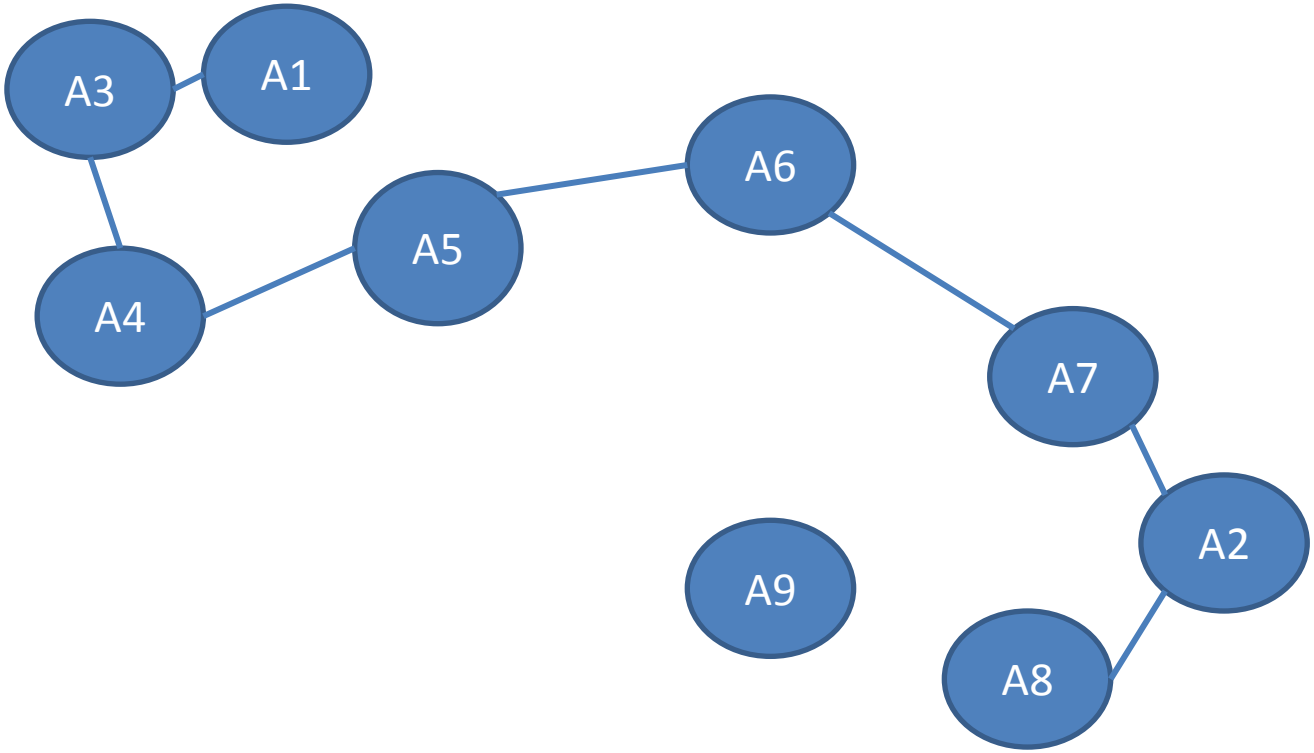
Алгоритмы образования комплексов различных типов

2.Цепной *кластер.*



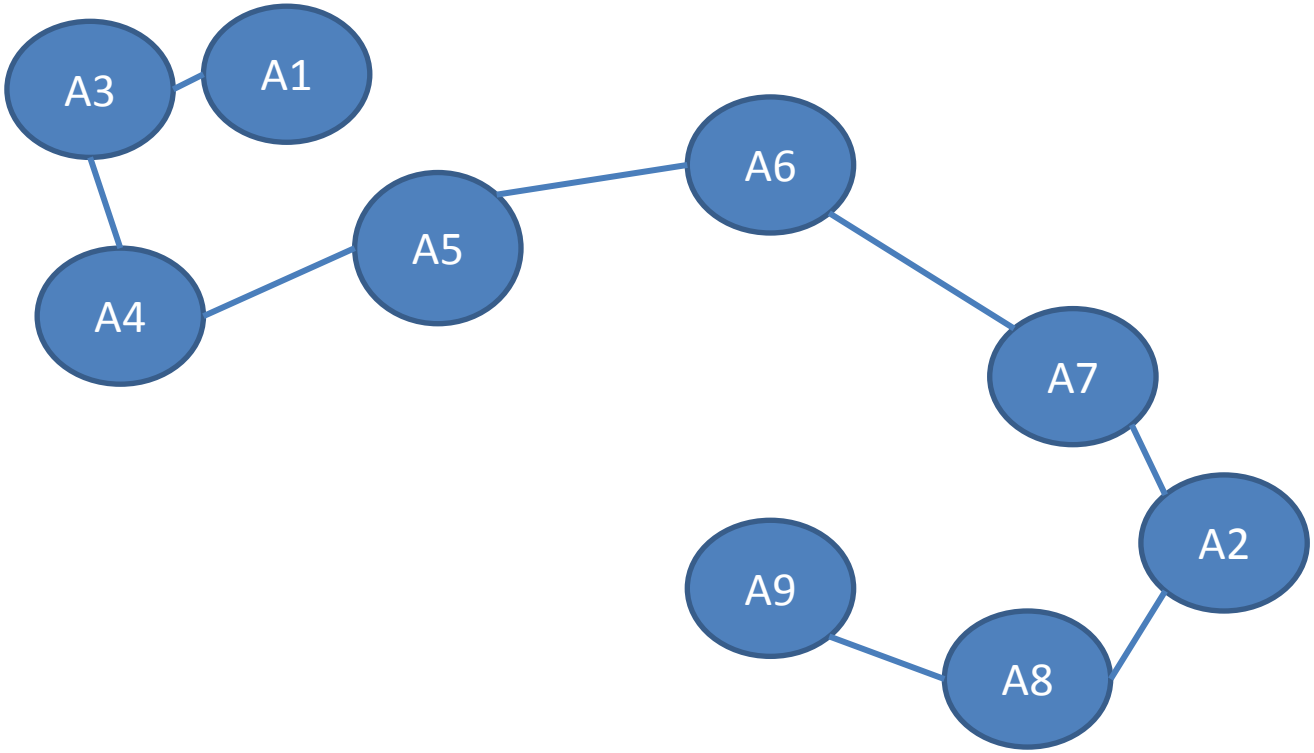
Алгоритмы образования комплексов различных типов

2.Цепной *кластер.*



Алгоритмы образования комплексов различных типов

2.Цепной *кластер.*



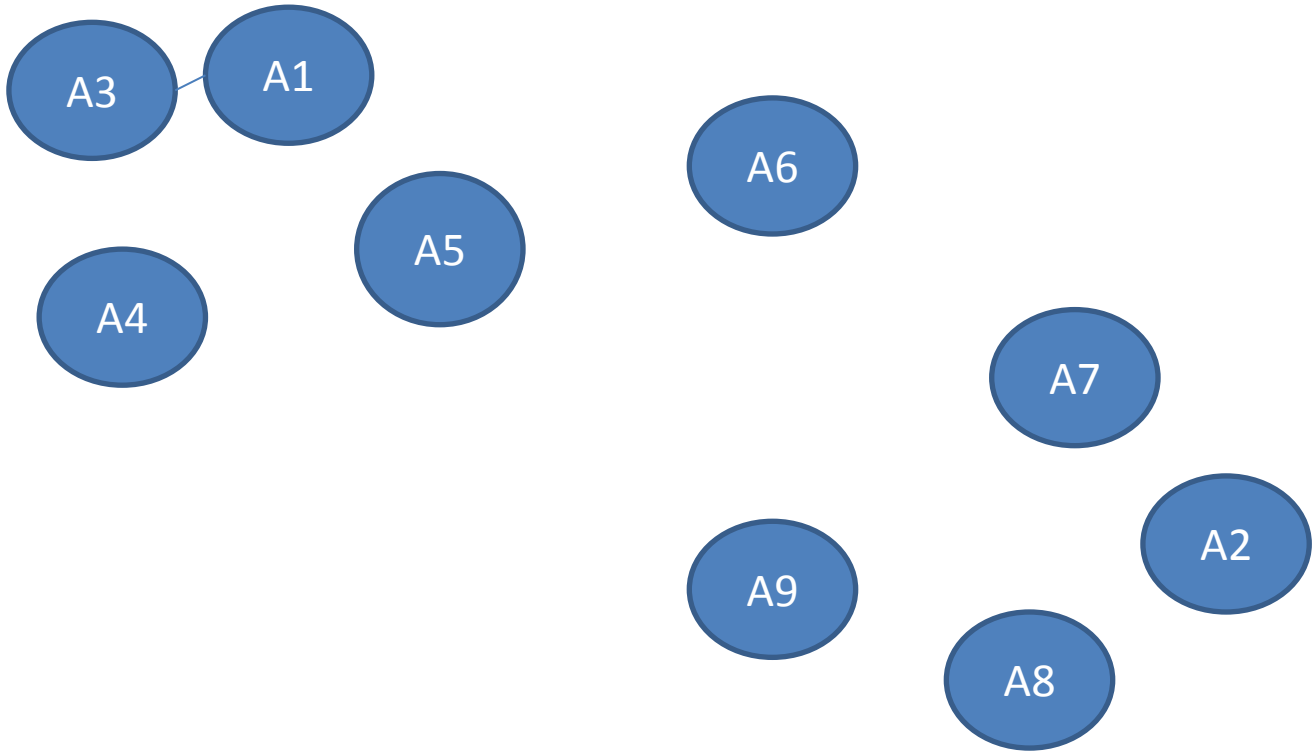
- если на каком-либо шаге построения цепного кластера минимальная величина будет не у одной, а у двух или более пар элементов, то в этом случае может быть построено несколько эквивалентных цепных кластеров.

3. Ассоциативно - цепной *кластер.*

- ассоциативный выявляет все элементы, ближайшие к ядру по различным свойствам,
- а цепной показывает связь данного начального элемента последовательно со всеми остальными элементами множества.

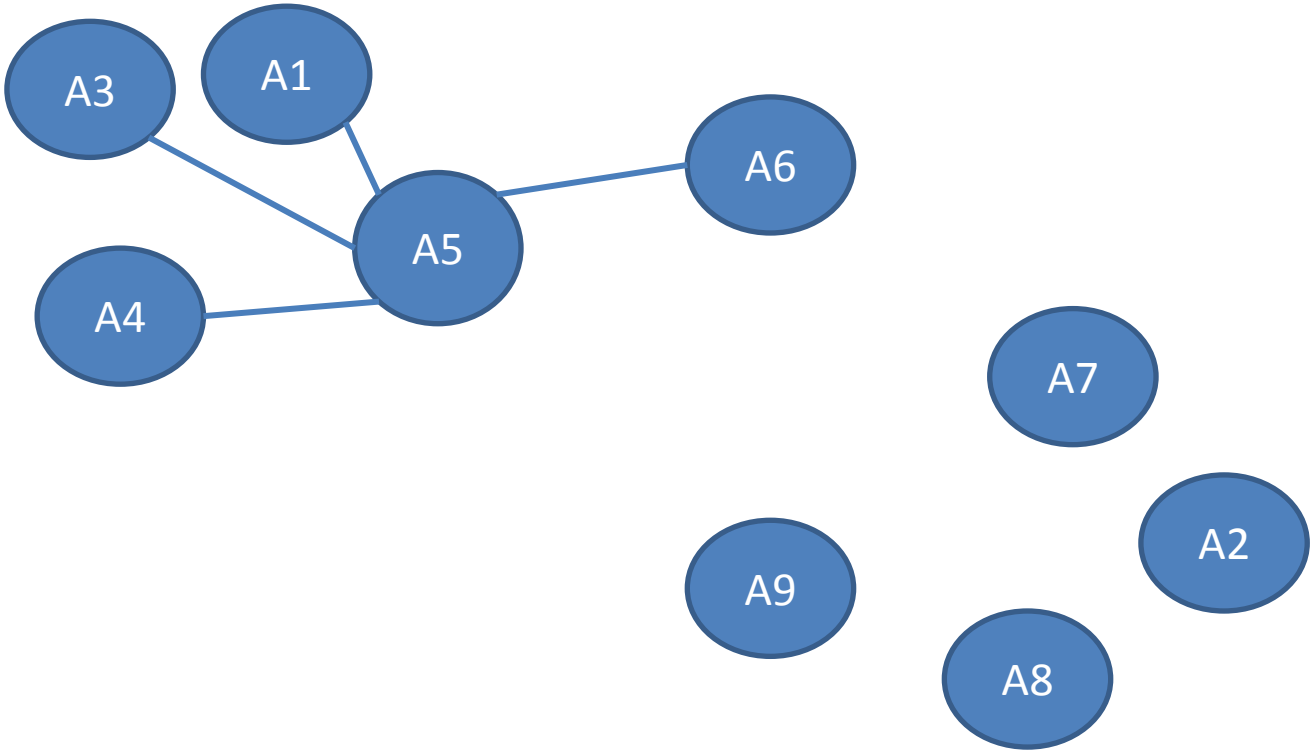
Алгоритмы образования комплексов различных типов

3. Ассоциативно - цепной *кластер.*



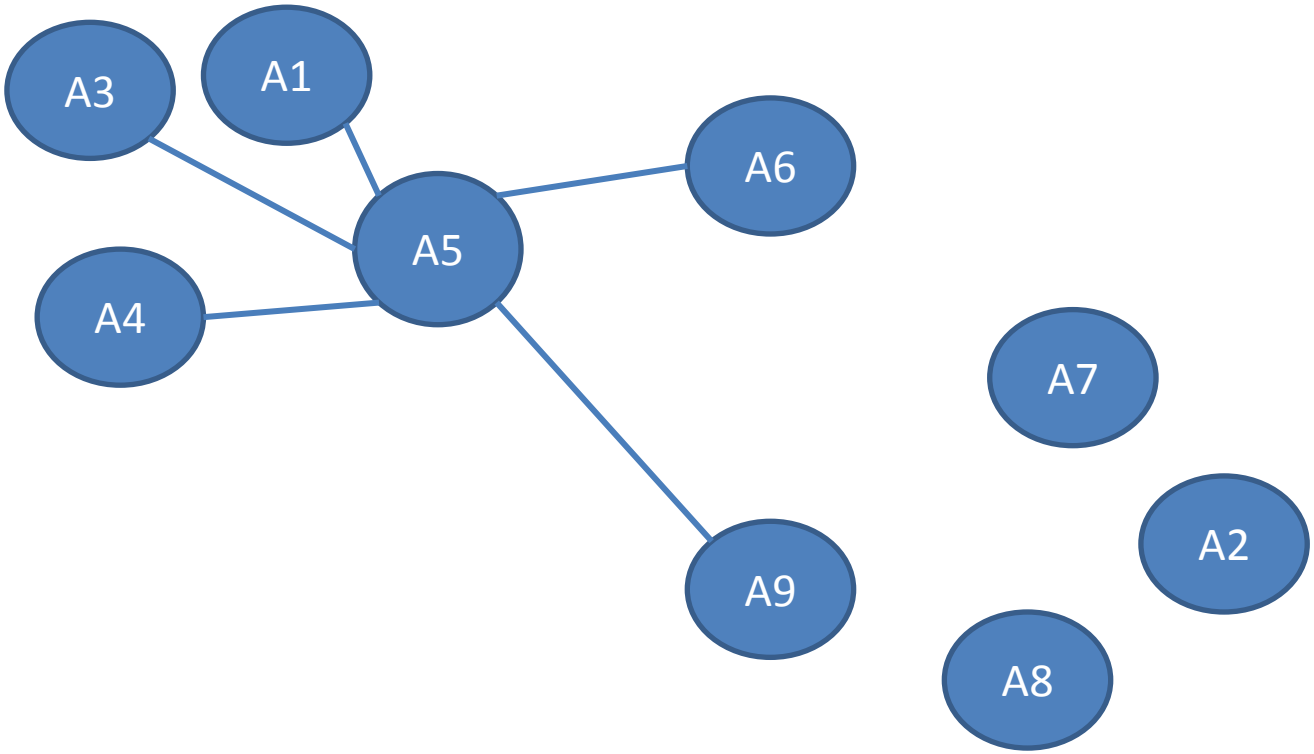
Алгоритмы образования комплексов различных типов

3. Ассоциативно - цепной *кластер.*



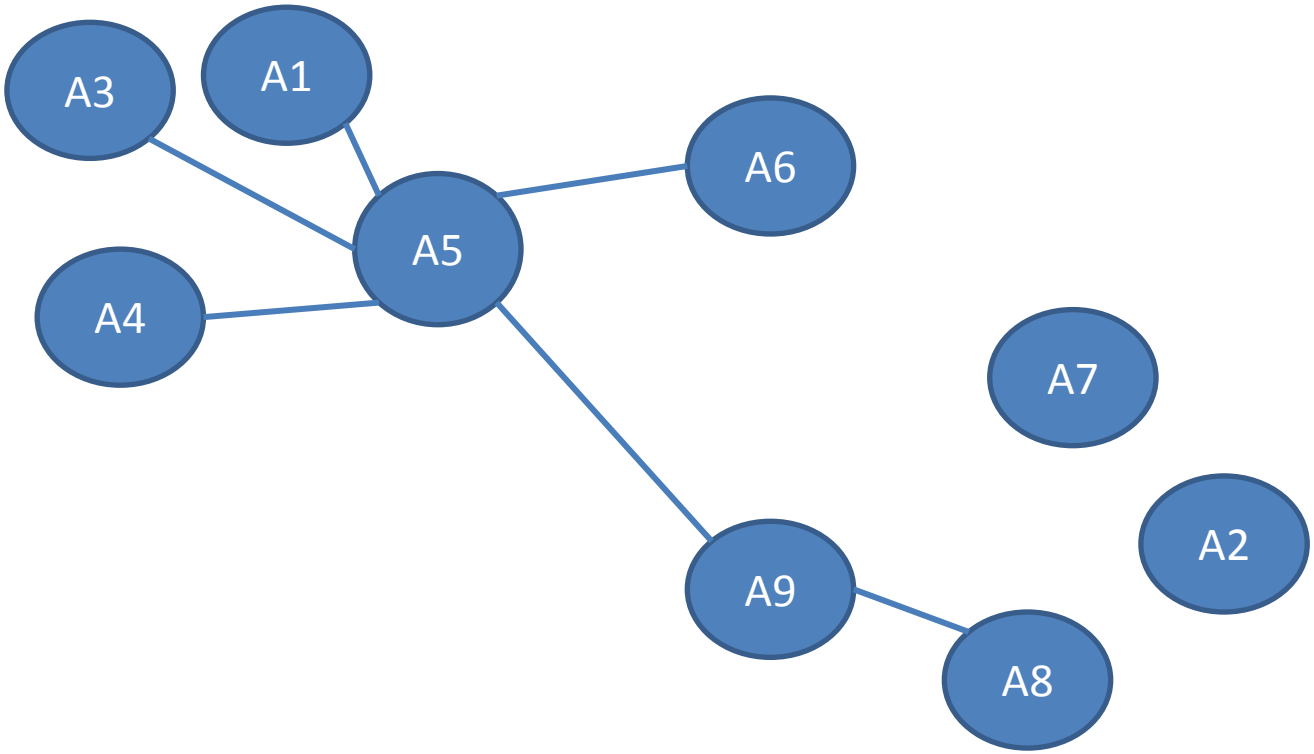
Алгоритмы образования комплексов различных типов

3. Ассоциативно - цепной *кластер.*



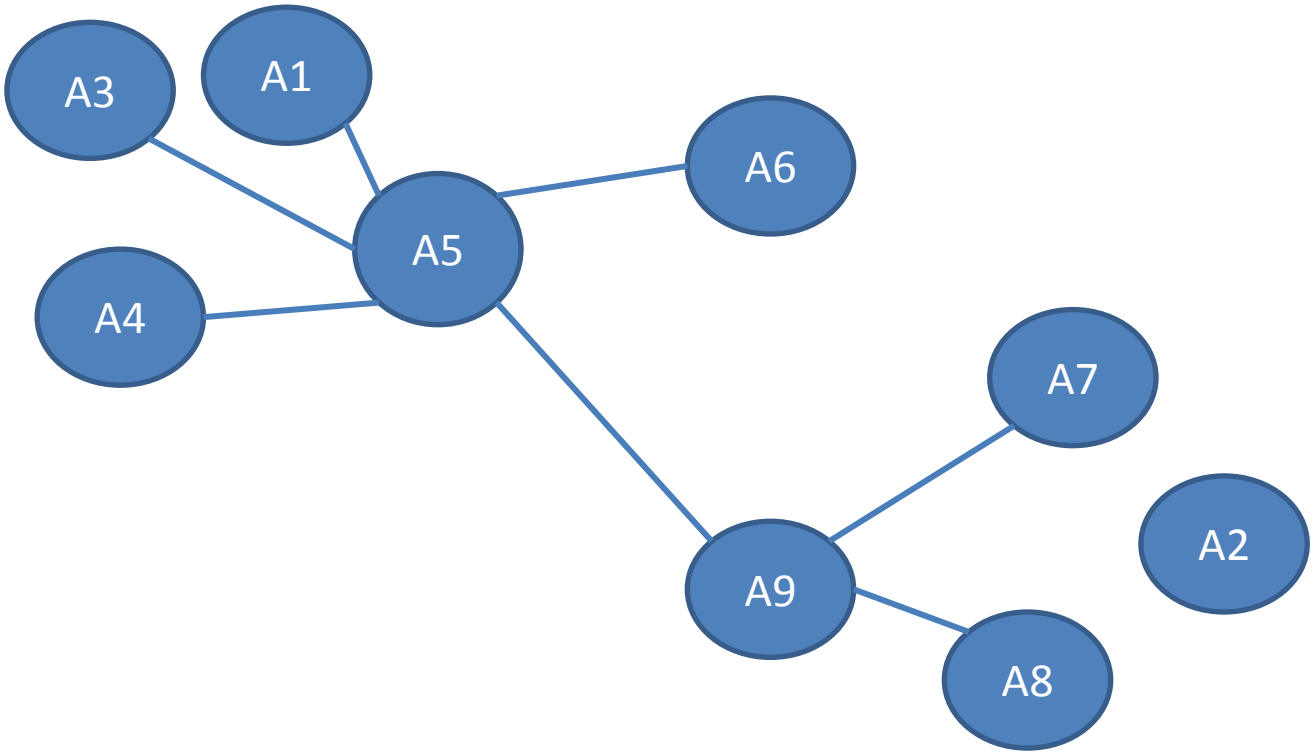
Алгоритмы образования комплексов различных типов

3. Ассоциативно - цепной *кластер.*



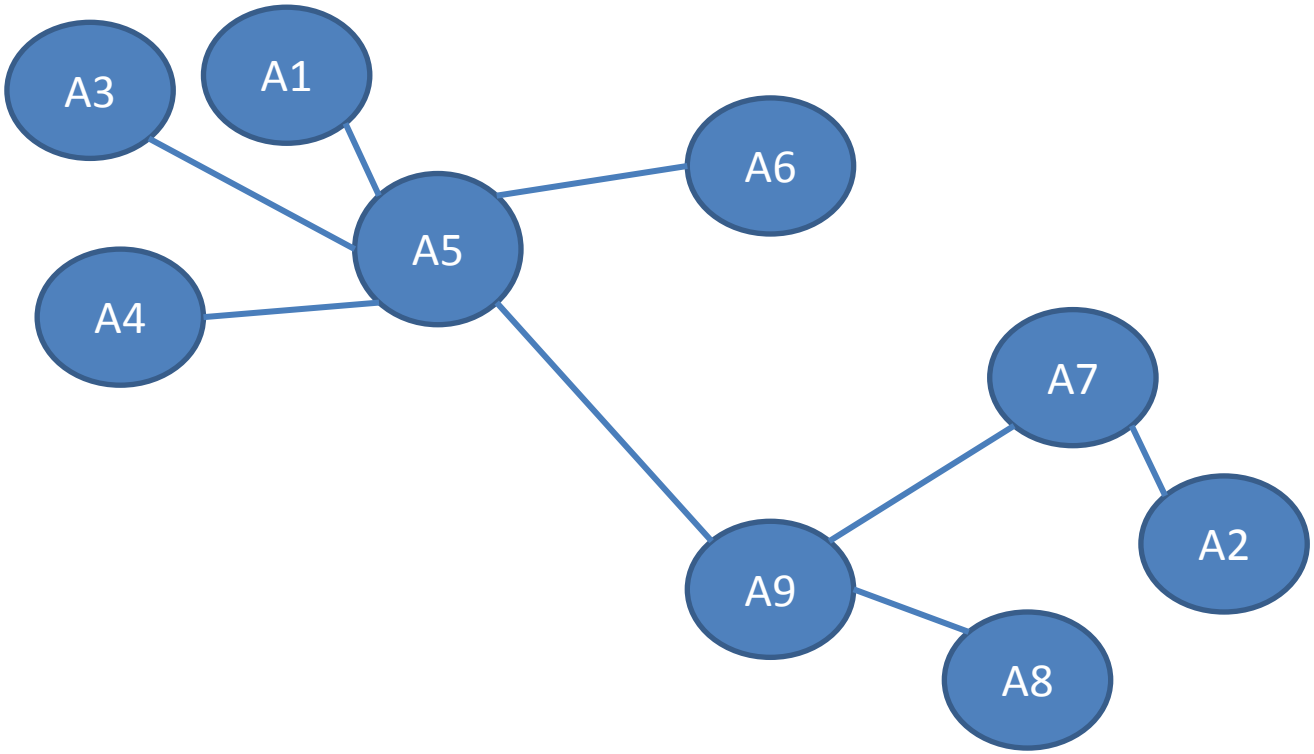
Алгоритмы образования комплексов различных типов

3. Ассоциативно - цепной *кластер.*



Алгоритмы образования комплексов различных типов

3. Ассоциативно - цепной *кластер.*

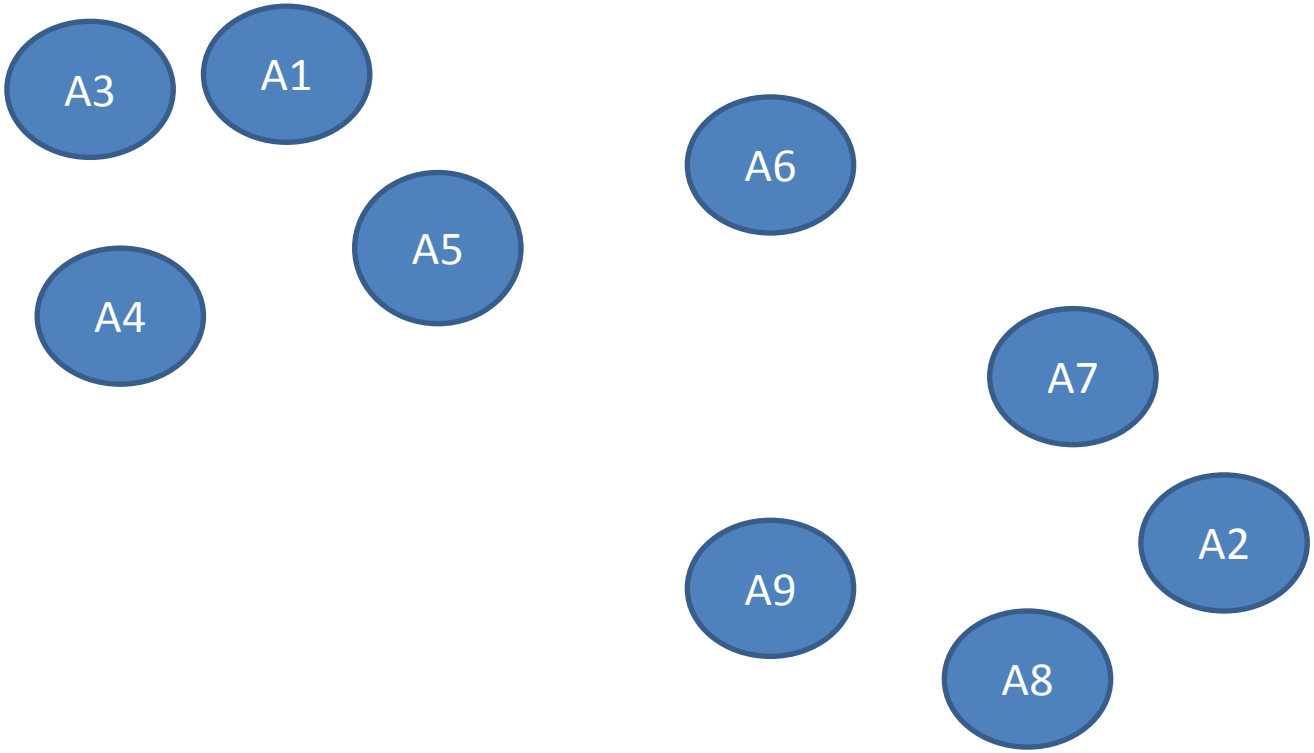


4. *Кластер-коллекция*.

- «***больше всего напоминают то, что принято называть коллекциями. Здесь различные неконкретные предметы объединяются на основе взаимного дополнения по какому-либо одному признаку и образуют единое целое, состоящее из разнородных, взаимно дополняющих друг друга частей***». И далее:
- «***Эта форма мышления часто соединяется с описанной выше ассоциативной формой. Тогда получается коллекция, составленная на основе различных признаков***»

Алгоритмы образования комплексов различных типов

4. Кластер-коллекция



- в результате применения алгоритма построения кластера-коллекции мы должны получить набор элементов, отличающихся друг от друга хотя бы по одному признаку.
- следующий алгоритм: сначала задается некоторый порог различия (или расстояния), при котором два элемента с разницей больше выбранного порога считаются различными. Очевидно, что результат (кластер-коллекция) будет зависеть от величины порога.

- Для первой матрицы – три кластера:

$$\{A_1 A_2 A_3 A_4 A_5 A_6 A_7\}\ \{A_8\}\ \{A_9\}.$$

- Для второй – четыре кластера:

$$\{A_1 A_2 A_3 A_4 A_5 A_6\}\ \{A_7\}\ \{A_8\}\ \{A_9\}.$$

- Для третьей – четыре кластера:

$$\{A_1 A_3 A_4 A_5 A_6\}\ \{A_2 A_7\}\ \{A_8\}\ \{A_9\}.$$

- Выбирая в соответствии с описанной выше процедурой пересечения и разности всех полученных кластеров, получим в результате следующий набор множеств:

- Выбирая в соответствии с описанной выше процедурой пересечения и разности всех полученных кластеров, получим в результате следующий набор множеств:

$$\{A_1 A_2 A_3 A_4 A_5 A_6 A_7\}\ \{A_8\}\ \{A_9\}.$$

$$\{A_1 A_2 A_3 A_4 A_5 A_6\}\ \{A_7\}\ \{A_8\}\ \{A_9\}.$$

$$\{A_1 A_3 A_4 A_5 A_6\}\ \{A_2 A_7\}\ \{A_8\}\ \{A_9\}.$$

- Выбирая в соответствии с описанной выше процедурой пересечения и разности всех полученных кластеров, получим в результате следующий набор множеств:

$$\{A_1 A_2 A_3 A_4 A_5 A_6 A_7\}\ \{A_8\}\ \{A_9\}.$$

$$\{A_1 A_2 A_3 A_4 A_5 A_6\}\ \{A_7\}\ \{A_8\}\ \{A_9\}.$$

$$\{A_1 A_3 A_4 A_5 A_6\}\ \{A_2 A_7\}\ \{A_8\}\ \{A_9\}.$$

$$\{A_1 A_3 A_4 A_5 A_6\}\ \{A_2\}\ \{A_7\}\ \{A_8\}\ \{A_9\}$$

- Выбирая в соответствии с описанной выше процедурой пересечения и разности всех полученных кластеров, получим в результате следующий набор множеств:

$$\{A_1 A_2 A_3 A_4 A_5 A_6 A_7\}\ \{A_8\}\ \{A_9\}.$$

$$\{A_1 A_2 A_3 A_4 A_5 A_6\}\ \{A_7\}\ \{A_8\}\ \{A_9\}.$$

$$\{A_1 A_3 A_4 A_5 A_6\}\ \{A_2 A_7\}\ \{A_8\}\ \{A_9\}.$$

$$\{A_1 A_3 A_4 A_5 A_6\}\ \{A_2\}\ \{A_7\}\ \{A_8\}\ \{A_9\}$$

- Таким образом, в кластер-коллекцию входят элементы A_8, A_9, A_7, A_2, и еще один (любой) элемент первого множества

- Шаг 1. Определить типы всех измерительных шкал, примененных для получения выборки эмпирических данных. Ответить на следующие вопросы: Применяются ли интервальные, порядковые, номинальные, дихотомические шкалы? Все ли используемые шкалы однотипны, или имеет место ситуация применения смешанных шкал?

- Шаг 2. Опираясь на исследовательский опыт, наметить план процедуры кластеризации, в зависимости от которого выбрать подходящий статистический пакет анализа данных, содержащего намеченный метод кластерного анализа.

- Шаг 3. Запустить пакет и ввести эмпирические данные в предлагаемую таблицу исходных данных, задав соответствующие названия и другие параметры переменных и сформировав, тем самым, матрицу «объект-признак».

- Шаг 4. В представленном в пакете блоке кластерного анализа последовательно выбрать направление кластеризации, меру сходства или различия для построения метрического пространства данных, глобальную стратегию кластеризации, адекватный конкретный метод кластерного анализа.

- Шаг 5. Выполнить запланированную и подготовленную процедуру кластеризации. Провести анализ и психологическую интерпретацию полученных результатов, осуществить дополнительную проверку их принципиальной правильности с использованием других методов кластеризации, другого статистического пакета и т.д.