Построение оптимальных композиций алгоритмов кластерного анализа разнородных данных

В.В.Бериков

План доклада

- Задача кластерного анализа

План доклада

- Задача кластерного анализа
- Коллективный подход в кластерном анализе

План доклада

- Задача кластерного анализа
- Коллективный подход в кластерном анализе
- Композиции алгоритмов с весами

План доклада

- Задача кластерного анализа
- Коллективный подход в кластерном анализе
- Композиции алгоритмов с весами
- Выбор оптимальных весов: модель ансамбля, критерий и алгоритм

План доклада

- Задача кластерного анализа
- Коллективный подход в кластерном анализе
- Композиции алгоритмов с весами
- Выбор оптимальных весов: модель ансамбля, критерий и алгоритм
- Численный эксперимент

Задача кластерного анализа

$A = \{a_1, ..., a_N\}$ – множество объектов.

Задача кластерного анализа

$A = \{a_1, ..., a_N\}$ – множество объектов.

1 вариант постановки: $X_1, ..., X_d$ – набор переменных, таблица данных $\mathbf{X} = \{x_{i,j}\}$, $x_{i,j} = X_j(a_i)$, $j = 1, ..., d$, $i = 1, ..., N$.

Задача кластерного анализа

$A = \{a_1, ..., a_N\}$ – множество объектов.

1 вариант постановки: $X_1, ..., X_d$ – набор переменных, таблица данных $\mathbf{X} = \{x_{i,j}\}$, $x_{i,j} = X_j(a_i)$, $j = 1, ..., d$, $i = 1, ..., N$.

2 вариант постановки: задана матрица попарных расстояний между объектами $(\rho_{i,j})$

Задача кластерного анализа

$A = \{a_1, ..., a_N\}$ – множество объектов.

1 вариант постановки: $X_1, ..., X_d$ – набор переменных, таблица данных $\mathbf{X} = \{x_{i,j}\}$, $x_{i,j} = X_j(a_i)$, $j = 1, ..., d$, $i = 1, ..., N$.

2 вариант постановки: задана матрица попарных расстояний между объектами $(\rho_{i,j})$

Требуется найти разбиение множества A на $K << N$ однородных классов (групп, кластеров) наилучшее по заданному критерию Q качества.

Задача кластерного анализа

$A = \{a_1, ..., a_N\}$ – множество объектов.

1 вариант постановки: $X_1, ..., X_d$ – набор переменных, таблица данных $\mathbf{X} = \{x_{i,j}\}$, $x_{i,j} = X_j(a_i)$, $j = 1, ..., d$, $i = 1, ..., N$.

2 вариант постановки: задана матрица попарных расстояний между объектами $(\rho_{i,j})$

Требуется найти разбиение множества A на $K << N$ однородных классов (групп, кластеров) наилучшее по заданному критерию Q качества.

Число кластеров может быть как выбрано заранее, так и не задано (нужно найти оптимальное K).

Применение

- Data Mining;
- Анализ изображений;
- Биоинформатика;
- Социологические исследования;
- и т.д.

Способы понимания однородности

- тривиальный: имеется некоторый признак, по которому проводится группирование;

Способы понимания однородности

- тривиальный: имеется некоторый признак, по которому проводится группирование;
- объекты имеют многомерное описание в пространстве характеристик; критерий однородности - функционал, зависящий от расстояний между объектами внутри групп и от расстояний между группами;

Способы понимания однородности

- тривиальный: имеется некоторый признак, по которому проводится группирование;
- объекты имеют многомерное описание в пространстве характеристик; критерий однородности - функционал, зависящий от расстояний между объектами внутри групп и от расстояний между группами;
- каждая группа соответствует своей модели возникновения данных; требуется оценить параметры модели и найти вариант группировки, наиболее точно соответствующий данным;

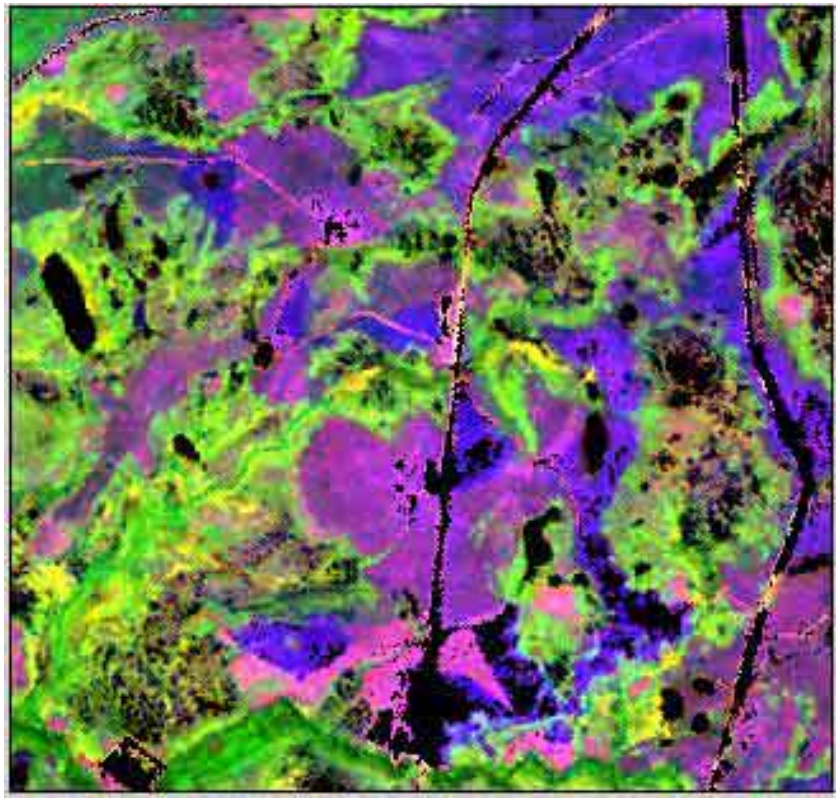
Способы понимания однородности

- тривиальный: имеется некоторый признак, по которому проводится группирование;
- объекты имеют многомерное описание в пространстве характеристик; критерий однородности - функционал, зависящий от расстояний между объектами внутри групп и от расстояний между группами;
- каждая группа соответствует своей модели возникновения данных; требуется оценить параметры модели и найти вариант группировки, наиболее точно соответствующий данным;
- кластер понимается как «сгусток», т.е. связная область по возможности меньшего объёма, содержащая как можно больше объектов. Вычисление объема основано на некоторой мере множества.

Способы понимания однородности

- тривиальный: имеется некоторый признак, по которому проводится группирование;
- объекты имеют многомерное описание в пространстве характеристик; критерий однородности - функционал, зависящий от расстояний между объектами внутри групп и от расстояний между группами;
- каждая группа соответствует своей модели возникновения данных; требуется оценить параметры модели и найти вариант группировки, наиболее точно соответствующий данным;
- кластер понимается как «сгусток», т.е. связная область по возможности меньшего объёма, содержащая как можно больше объектов. Вычисление объема основано на некоторой мере множества.
- другие способы, использующие конкретику прикладной области.

Пример: сегментация спутникового изображения

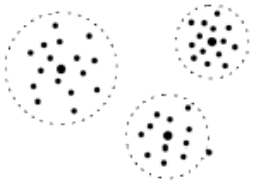


Требуется выделить участки, соответствующие различным типам растительности (смешанный лес; участок после пожара; болото; пойма реки и т.д.)

Кластерные структуры

ленточные кластеры

кластеры могут образовываться не по сходству, а по иным типам регулярностей



кластеры с центром

кластеры могут вообще отсутствовать

кластеры могут соединяться перемычками

кластеры могут накладываться на разреженный фон из редко расположенных объектов

кластеры могут перекрываться

Основные этапы кластерного анализа

- Формирование системы переменных (feature selection, feature extraction), нормировка, задание числа кластеров;

Основные этапы кластерного анализа

- Формирование системы переменных (feature selection, feature extraction), нормировка, задание числа кластеров;

- Выбор критерия качества, задание расстояний между объектами и группами объектов;

Основные этапы кластерного анализа

- Формирование системы переменных (feature selection, feature extraction), нормировка, задание числа кластеров;
- Выбор критерия качества, задание расстояний между объектами и группами объектов;
- Группировка объектов с помощью некоторого алгоритма перебора вариантов разбиения на группы;

Основные этапы кластерного анализа

- Формирование системы переменных (feature selection, feature extraction), нормировка, задание числа кластеров;
- Выбор критерия качества, задание расстояний между объектами и группами объектов;
- Группировка объектов с помощью некоторого алгоритма перебора вариантов разбиения на группы;

трудоемкость!!!

- Представление результатов («типичный» объект; «центроид», таксон).

- Представление результатов («типичный» объект; «центроид», таксон).

  Таксон - подобласть пространства переменных минимального объема, имеющую некоторую заданную форму и содержащую точки соответствующей группы.

- Представление результатов («типичный» объект; «центроид», таксон).

  Таксон - подобласть пространства переменных минимального объема, имеющую некоторую заданную форму и содержащую точки соответствующей группы.

- Определение качества полученной группировки

  - содержательная интерпретация;

- Представление результатов («типичный» объект; «центроид», таксон).

  Таксон - подобласть пространства переменных минимального объема, имеющую некоторую заданную форму и содержащую точки соответствующей группы.

- Определение качества полученной группировки

  - содержательная интерпретация;
  - статистические критерии

    (проверка статистических гипотез о случайном образовании кластеров, их однородности и т.д.);

- Представление результатов («типичный» объект; «центроид», таксон).

  Таксон - подобласть пространства переменных минимального объема, имеющую некоторую заданную форму и содержащую точки соответствующей группы.

- Определение качества полученной группировки

  - содержательная интерпретация;
  - статистические критерии

    (проверка статистических гипотез о случайном образовании кластеров, их однородности и т.д.);
  - индексы качества.

Индексы качества (validity indexes): «внешние» и «внутренние»
Внешние – сравнение с «истинной» группировкой

Индексы качества (validity indexes): «внешние» и «внутренние»

Внешние – сравнение с «истинной» группировкой

проблема – перестановки номеров кластеров эквивалентны

Индексы качества (validity indexes): «внешние» и «внутренние»

Внешние – сравнение с «истинной» группировкой

проблема – перестановки номеров кластеров эквивалентны

Индекс Ранда $RI = \frac{A + D}{C_N^2}$, где A - число пар объектов, которые входят в одни и те же группы, D - число пар, которые входят в разные группы в сравниваемых вариантах.

Индексы качества (validity indexes): «внешние» и «внутренние»

Внешние – сравнение с «истинной» группировкой

проблема – перестановки номеров кластеров эквивалентны

Индекс Ранда $RI = \frac{A + D}{C_N^2}$, где A - число пар объектов, которые входят в одни и те же группы, D - число пар, которые входят в разные группы в сравниваемых вариантах.

«Исправленный» индекс Ранда (adjusted Rand index)

$$ARI = \frac{RI - E(RI)}{RI_{\max} - E(RI)}, \quad \text{или}$$

$$ARI = \frac{\sum_{k,l} C_{N_{k,l}}^2 - Q_1 Q_2 \big/ C_N^2}{\frac{1}{2}(Q_1 + Q_2) - Q_1 Q_2 \big/ C_N^2}, \text{ где } Q_1 = \sum_k C_{N_1^{(k)}}^2, \; Q_2 = \sum_k C_{N_2^{(l)}}^2.$$

Индекс нормированной взаимной информации

Пусть $X \in \{1,\ldots,K_1\}$, $Y \in \{1,\ldots,K_2\}$,

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}},$$

где $I(X,Y)$ – количество взаимной информации между X и Y, $H(X)$ и $H(Y)$ – энтропия.

Индекс нормированной взаимной информации

Пусть $X \in \{1,\ldots,K_1\}$, $Y \in \{1,\ldots,K_2\}$,

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}},$$

где $I(X,Y)$ – количество взаимной информации между X и Y, $H(X)$ и $H(Y)$ – энтропия. Выборочная оценка:

$$\varphi^{(\mathrm{NMI})} = \frac{\sum_{k,l} N_{k,l} \log\left(\frac{N \cdot N_{k,l}}{N_1^{(k)} N_2^{(l)}}\right)}{\sqrt{\left(\sum_k N_1^{(k)} \log\frac{N_1^{(k)}}{N}\right)\left(\sum_l N_2^{(l)} \log\frac{N_2^{(l)}}{N}\right)}}.$$

Индекс нормированной взаимной информации

Пусть $X \in \{1,\ldots,K_1\}$, $Y \in \{1,\ldots,K_2\}$,

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}},$$

где $I(X,Y)$ – количество взаимной информации между X и Y, $H(X)$ и $H(Y)$ – энтропия. Выборочная оценка:

$$\varphi^{(\mathrm{NMI})} = \frac{\sum_{k,l} N_{k,l} \log\left(\frac{N \cdot N_{k,l}}{N_1^{(k)} N_2^{(l)}}\right)}{\sqrt{\left(\sum_k N_1^{(k)} \log\frac{N_1^{(k)}}{N}\right)\left(\sum_l N_2^{(l)} \log\frac{N_2^{(l)}}{N}\right)}}.$$

$\varphi^{(\mathrm{NMI})} = 0$: ($X$ и Y независимы).

Индекс нормированной взаимной информации

Пусть $X \in \{1,\ldots,K_1\}$, $Y \in \{1,\ldots,K_2\}$,

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}},$$

где $I(X,Y)$ – количество взаимной информации между X и Y, $H(X)$ и $H(Y)$ – энтропия. Выборочная оценка:

$$\varphi^{(\mathrm{NMI})} = \frac{\sum_{k,l} N_{k,l} \log\left(\frac{N \cdot N_{k,l}}{N_1^{(k)} N_2^{(l)}}\right)}{\sqrt{\left(\sum_k N_1^{(k)} \log\frac{N_1^{(k)}}{N}\right)\left(\sum_l N_2^{(l)} \log\frac{N_2^{(l)}}{N}\right)}}.$$

$\varphi^{(\mathrm{NMI})} = 0$: ($X$ и Y независимы).

$\varphi^{(\mathrm{NMI})} = 1$: варианты группировки совпадают.

«Внутренние» индексы – агрегирование мер близости объектов в группах и мер взаимной удаленности групп

«Внутренние» индексы – агрегирование мер близости объектов в группах и мер взаимной удаленности групп

Пример: индекс Дунна (Dunn):

$$D(C) = \frac{\min\limits_{C_k \in C} \{ \min\limits_{C_l \in C \setminus C_k} \{\delta(C_k, C_l)\}\}}{\max\limits_{C_k \in C} \{\Delta(C_k)\}},$$

где $\delta(C_k, C_l) = \min\limits_{o^{(i)} \in C_k} \min\limits_{o^{(j)} \in C_l} \{\rho_e(x^{(i)}, x^{(j)})\}$, $\Delta(C_k) = \max\limits_{o^{(i)}, o^{(j)} \in C_k} \{\rho_e(o^{(i)}, o^{(j)})\}$.

«Внутренние» индексы – агрегирование мер близости объектов в группах и мер взаимной удаленности групп

Пример: индекс Дунна (Dunn):

$$D(C) = \frac{\min\limits_{C_k \in C} \{ \min\limits_{C_l \in C \setminus C_k} \{\delta(C_k, C_l)\}\}}{\max\limits_{C_k \in C} \{\Delta(C_k)\}},$$

где $\delta(C_k, C_l) = \min\limits_{o^{(i)} \in C_k} \min\limits_{o^{(j)} \in C_l} \{\rho_e(x^{(i)}, x^{(j)})\}$, $\Delta(C_k) = \max\limits_{o^{(i)}, o^{(j)} \in C_k} \{\rho_e(o^{(i)}, o^{(j)})\}$.

Степень удаленности кластеров оценивается как расстояние между ближайшими объектами, им принадлежащими, а компактность кластера - как его «диаметр».

Основные подходы в кластерном анализе

- Вероятностный (расщепление смеси распределений; непараметрическое восстановление плотности);

Основные подходы в кластерном анализе

- Вероятностный (расщепление смеси распределений; непараметрическое восстановление плотности);
- Использующий аналогии с «центром тяжести» (алгоритмы FOREL, k-средних);

Основные подходы в кластерном анализе

- Вероятностный (расщепление смеси распределений; непараметрическое восстановление плотности);
- Использующий аналогии с «центром тяжести» (алгоритмы FOREL, k-средних);
- Основанный на теории графов (алгоритм кратчайшего незамкнутого пути; дерево иерархической группировки);

Основные подходы в кластерном анализе

- Вероятностный (расщепление смеси распределений; непараметрическое восстановление плотности);
- Использующий аналогии с «центром тяжести» (алгоритмы FOREL, k-средних);
- Основанный на теории графов (алгоритм кратчайшего незамкнутого пути; дерево иерархической группировки);
- k ближайших соседей и его обобщения (при нахождении меры близости учитывается расположение других соседних точек);

Основные подходы в кластерном анализе

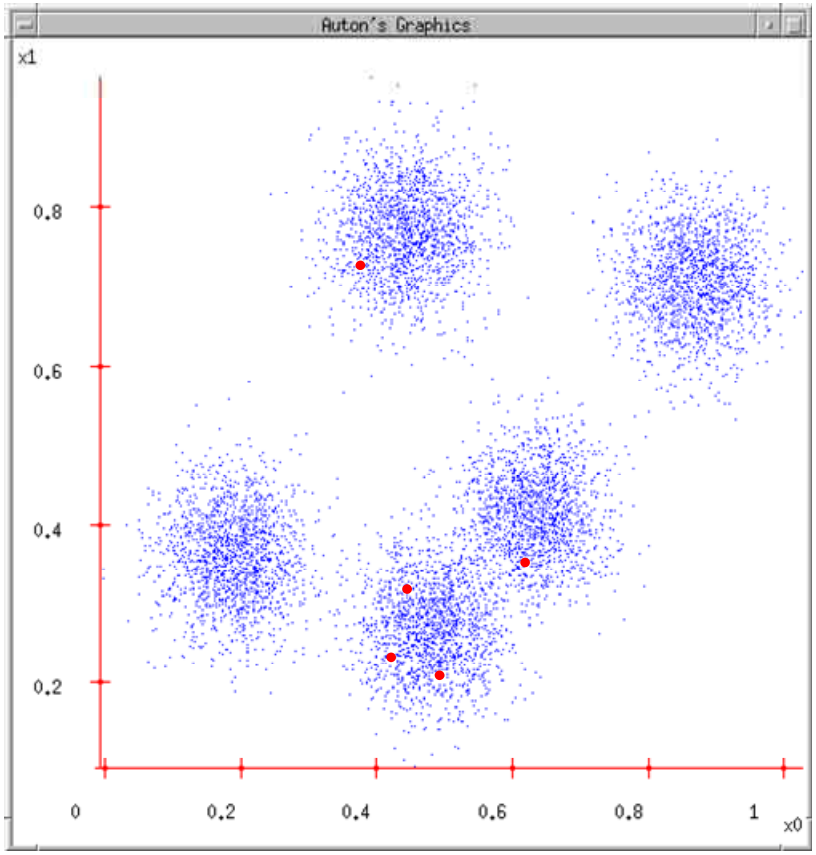
- Вероятностный (расщепление смеси распределений; непараметрическое восстановление плотности);
- Использующий аналогии с «центром тяжести» (алгоритмы FOREL, k-средних);
- Основанный на теории графов (алгоритм кратчайшего незамкнутого пути; дерево иерархической группировки);
- k ближайших соседей и его обобщения (при нахождении меры близости учитывается расположение других соседних точек);
- Нечеткие алгоритмы (Fuzzy C-means);

Основные подходы в кластерном анализе

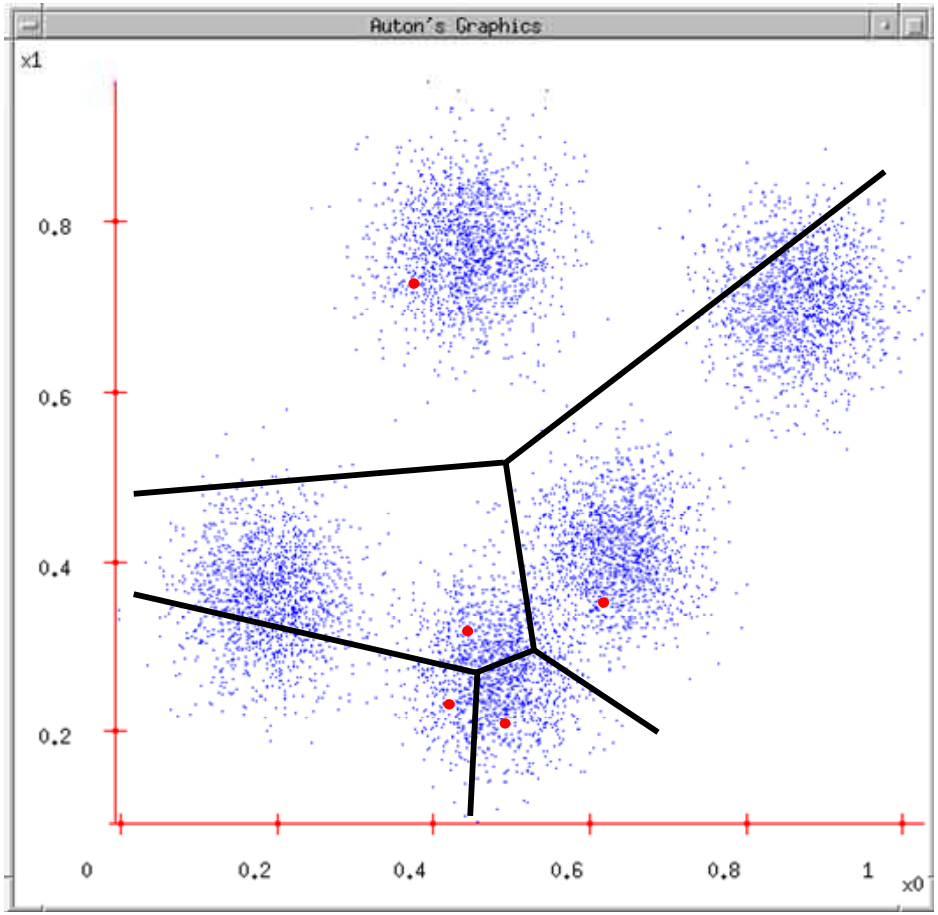
- Вероятностный (расщепление смеси распределений; непараметрическое восстановление плотности);
- Использующий аналогии с «центром тяжести» (алгоритмы FOREL, k-средних);
- Основанный на теории графов (алгоритм кратчайшего незамкнутого пути; дерево иерархической группировки);
- k ближайших соседей и его обобщения (при нахождении меры близости учитывается расположение других соседних точек);
- Нечеткие алгоритмы (Fuzzy C-means);
- Искусственные нейронные сети.

Алгоритм K-средних

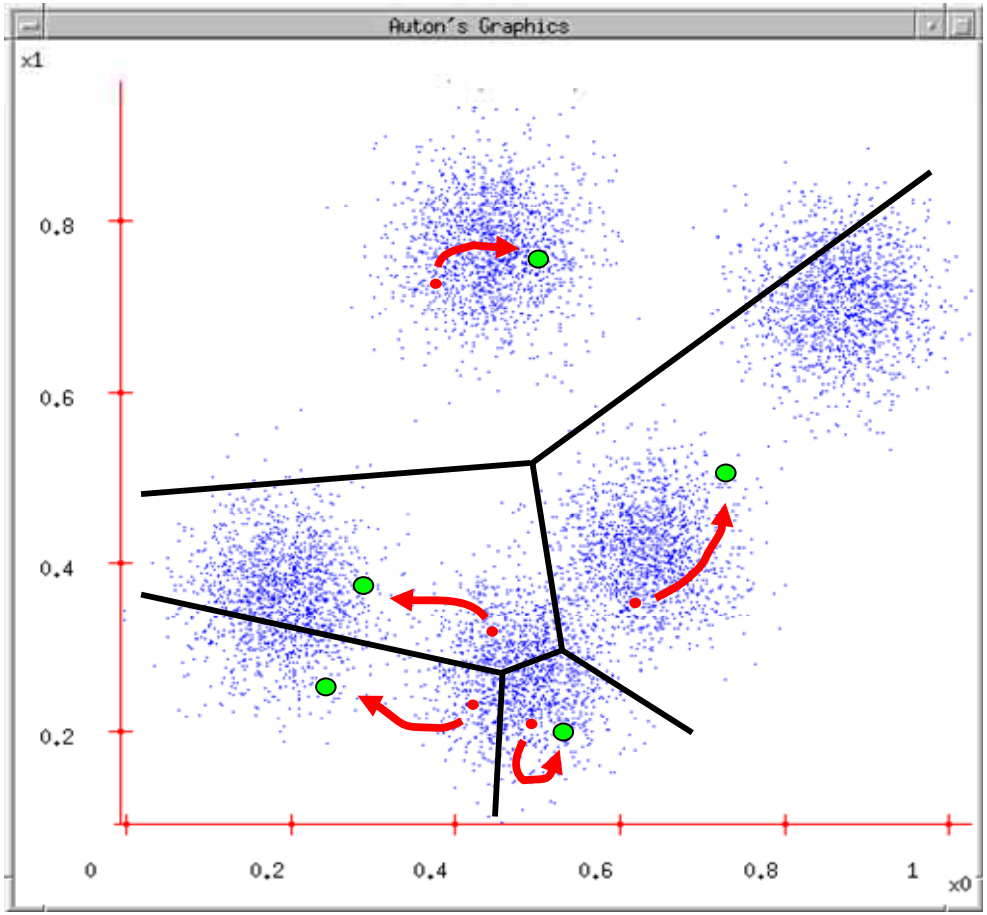
Пусть задано требуемое число кластеров K.
1. Произвольным образом задаются вектора начальных «центров притяжений».



2. Для каждого центра ищутся точки, которые являются наиболее близкими к этому центру.



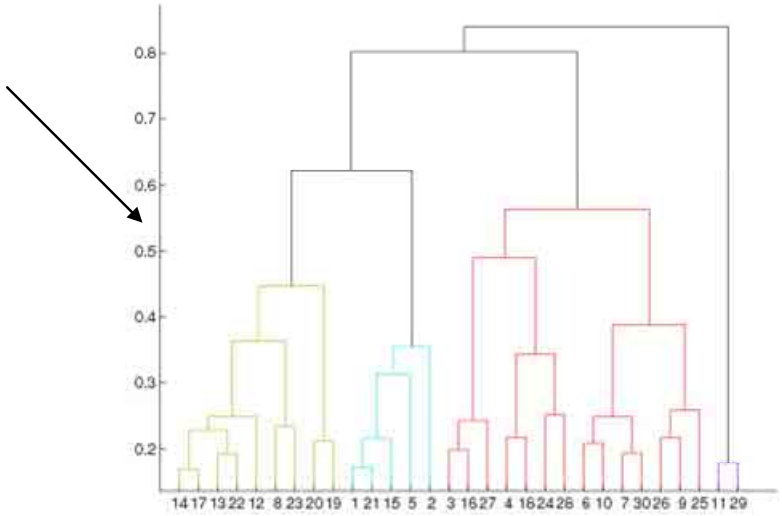
3. Из этих точек формируются новые группы; определяются их «центры тяжести» и т.д., пока группы не перестанут менять состав.



Иерархические методы кластерного анализа

Иерархия – расположение частей или элементов целого в порядке от высшего к низшему. Иерархическая группировка – такая группировка, в которой исходное множество объектов образуется из нескольких групп, каждая из которых также состоит из нескольких групп и т.д. Отношения между группами изображаются в виде дерева (дендрограмма)

расстояния между группами (точками)



Агломеративный алгоритм построения дендрограммы :

1. Найти ближайшую пару объектов (групп). При этом можно использовать расстояние по «ближайшему соседу», «дальнему соседу» и т.д.;

Агломеративный алгоритм построения дендрограммы :

1. Найти ближайшую пару объектов (групп). При этом можно использовать расстояние по «ближайшему соседу», «дальнему соседу» и т.д.;
2. Объединить найденную пару в одну группу;

Агломеративный алгоритм построения дендрограммы :

1. Найти ближайшую пару объектов (групп). При этом можно использовать расстояние по «ближайшему соседу», «дальнему соседу» и т.д.;
2. Объединить найденную пару в одну группу;
3. Найти следующую ближайшую пару;

Агломеративный алгоритм построения дендрограммы :

1. Найти ближайшую пару объектов (групп). При этом можно использовать расстояние по «ближайшему соседу», «дальнему соседу» и т.д.;
2. Объединить найденную пару в одну группу;
3. Найти следующую ближайшую пару;
4. Объединить найденную пару и т.д.,

пока все объекты будут объединены в одну группу.

Коллективный подход в кластерном анализе (Миркин Б.Г., Рязанов В.В., Торchy A., Jain A., Fred A., Ghosh J. и др.)

- Повышается устойчивость результатов по отношению к выбору параметров работы алгоритма;

Коллективный подход в кластерном анализе (Миркин Б.Г., Рязанов В.В., Торсhy A., Jain A., Fred A., Ghosh J. и др.)

- Повышается устойчивость результатов по отношению к выбору параметров работы алгоритма;
- Доказано, что при выполнении определенных «естественных» условий качество группировки улучшается;

Коллективный подход в кластерном анализе (Миркин Б.Г., Рязанов В.В., Торchy A., Jain A., Fred A., Ghosh J. и др.)

- Повышается устойчивость результатов по отношению к выбору параметров работы алгоритма;
- Доказано, что при выполнении определенных «естественных» условий качество группировки улучшается;
- Группировка с «разных точек зрения»:
  - различные подмножества переменных;
  - различные подмножества объектов;

Коллективный подход в кластерном анализе (Миркин Б.Г., Рязанов В.В., Торchy A., Jain A., Fred A., Ghosh J. и др.)

- Повышается устойчивость результатов по отношению к выбору параметров работы алгоритма;
- Доказано, что при выполнении определенных «естественных» условий качество группировки улучшается;
- Группировка с «разных точек зрения»:
  - различные подмножества переменных;
  - различные подмножества объектов;
- Возможность проведения распределенных вычислений (при различном местоположении подмножеств объектов или переменных).

Два основных направления в построении композиций (ансамблей) алгоритмов

1. Нахождение консенсусного разбиения

Два основных направления в построении композиций (ансамблей) алгоритмов

1. Нахождение консенсусного разбиения

Пусть имеется набор разбиений $\{P_1, ..., P_L\}$ множества объектов на группы

Два основных направления в построении композиций (ансамблей) алгоритмов

1. Нахождение консенсусного разбиения

Пусть имеется набор разбиений $\{P_1, ..., P_L\}$ множества объектов на группы

Требуется найти согласованное (консенсусное) разбиение P^*, оптимальное по некоторому заданному критерию.

$$P^* = arg \max_{P \in \mathbf{P}} \sum_{l=1}^{L} w_l \, \delta(P, P_l),$$

где δ - мера близости пары разбиений, $\mathbf{P}$ - множество всех разбиений, $w_l \geq 0$ - вес of l-го разбиения, $l = 1, ..., L$; $\sum_l w_l = 1$.

Два основных направления в построении композиций (ансамблей) алгоритмов

1. Нахождение консенсусного разбиения

Пусть имеется набор разбиений $\{P_1, ..., P_L\}$ множества объектов на группы

Требуется найти согласованное (консенсусное) разбиение P^*, оптимальное по некоторому заданному критерию.

$$P^* = arg \max_{P \in \mathbf{P}} \sum_{l=1}^{L} w_l \, \delta(P, P_l),$$

где δ - мера близости пары разбиений, $\mathbf{P}$ - множество всех разбиений, $w_l \geq 0$ - вес of l-го разбиения, $l = 1, ..., L$; $\sum_l w_l = 1$.

Например, в соответствии с принципом максимизации количества взаимной информации нужно найти такое P^*, для которого суммарная мера $\varphi^{(NMI)}$ максимальна.

2. Использование согласованной матрицы различия объектов

I-й этап: l-й вариант разбиения $\rightarrow$ бинарная матрица различий

$$H_l = \{ h_l(i, j) \},$$

где $h_l(i, j) = 1$, если $o^{(i)}$ и $o^{(j)}$ принадлежат разным кластерам;

$h_l(i, j) = 0$, иначе, $\quad i, j = 1, 2, \ldots, N, \;\; i \neq j$.

2. Использование согласованной матрицы различия объектов

I-й этап: l-й вариант разбиения $\rightarrow$ бинарная матрица различий

$$H_l = \{ h_l(i,j) \},$$

где $h_l(i,j) = 1$, если $o^{(i)}$ и $o^{(j)}$ принадлежат разным кластерам;

$h_l(i,j) = 0$, иначе, $\quad i, j = 1, 2, \ldots, N, \;\; i \neq j$.

Согласованная матрица различий $H^* = \{ h^*(i,j) \}$,

$$h^*(i,j) = \sum_{l=1}^{L} w_l h_l(i,j).$$

2. Использование согласованной матрицы различия объектов

I-й этап: l-й вариант разбиения $\rightarrow$ бинарная матрица различий

$$H_l = \{ h_l(i, j) \},$$

где $h_l(i, j) = 1$, если $o^{(i)}$ и $o^{(j)}$ принадлежат разным кластерам;

$h_l(i, j) = 0$, иначе, $\quad i, j = 1, 2, \ldots, N, \;\; i \neq j$.

Согласованная матрица различий $H^* = \{ h^*(i, j) \}$,

$$h^*(i, j) = \sum_{l=1}^{L} w_l h_l(i, j).$$

II-й этап: Нахождение итогового варианта группировки: любой алгоритм, который в качестве входной информации использует расстояния между объектами (например, алгоритм построения дендрограммы).

Существуют различные варианты определения весов, например:

- равные веса ($w_l \equiv 1/L$);

Существуют различные варианты определения весов, например:

- равные веса ($w_l \equiv 1/L$);
- пропорциональны значениям индекса качества группировки;

Существуют различные варианты определения весов, например:

- равные веса ($w_l \equiv 1/L$);
- пропорциональны значениям индекса качества группировки;
- веса переменных обратно пропорциональны дисперсии проекций на координатные оси;

Существуют различные варианты определения весов, например:

- равные веса ($w_l \equiv 1/L$);
- пропорциональны значениям индекса качества группировки;
- веса переменных обратно пропорциональны дисперсии проекций на координатные оси;
- учитывается мера разнообразия вариантов: похожим вариантам разбиения приписывается меньший вес.

Выбор оптимальных весов на основе модели ансамбля

Пусть с помощью набора алгоритмов кластерного анализа $\mu_1,...,\mu_M$ строятся $L_1,...,L_M$ вариантов разбиения на кластеры;

Выбор оптимальных весов на основе модели ансамбля

Пусть с помощью набора алгоритмов кластерного анализа $\mu_1,...,\mu_M$ строятся $L_1,...,L_M$ вариантов разбиения на кластеры;

обозначим

$$\overline{h}(i,j) = \sum_{m=1}^{M} \alpha_m(i,j) \frac{1}{L_m} \sum_{l=1}^{L_m} h_m(i,j)$$

где $\alpha_m(i,j)$ - вес («компетентность» алгоритма для пары i,j)

Выбор оптимальных весов на основе модели ансамбля

Пусть с помощью набора алгоритмов кластерного анализа $\mu_1,...,\mu_M$ строятся $L_1,...,L_M$ вариантов разбиения на кластеры;

обозначим

$$\bar{h}(i,j) = \sum_{m=1}^{M} \alpha_m(i,j) \frac{1}{L_m} \sum_{l=1}^{L_m} h_m(i,j)$$

где $\alpha_m(i,j)$ - вес («компетентность» алгоритма для пары i,j)

Необходима модель, позволяющая оценить «компетентность» алгоритмов и связать ее с наблюдаемыми характеристиками ансамбля.

Модель попарной классификации с латентными классами

- Y - непосредственно ненаблюдаемая (латентная) *переменная* (номер класса), $Y \in \{1,...,K\}$;

Модель попарной классификации с латентными классами

- Y - непосредственно ненаблюдаемая (латентная) *переменная* (номер класса), $Y \in \{1,...,K\}$;
- *Априорные вероятности* $P_k = P(Y = k)$, $\sum_{k=1}^{K} P_k = 1$;

Модель попарной классификации с латентными классами

- Y - непосредственно ненаблюдаемая (латентная) *переменная* (номер класса), $Y \in \{1,...,K\}$;
- *Априорные вероятности* $P_k = P(Y = k)$, $\sum_{k=1}^{K} P_k = 1$;
- *Условные распределения* $p(x \mid Y = k) = p_k(x)$, $k = 1,...,K$.

Модель попарной классификации с латентными классами

- Y - непосредственно ненаблюдаемая (латентная) *переменная* (номер класса), $Y \in \{1,...,K\}$;
- *Априорные вероятности* $P_k = P(Y = k)$, $\sum_{k=1}^{K} P_k = 1$;
- *Условные распределения* $p(x \mid Y = k) = p_k(x)$, $k = 1,...,K$.
- Для объекта, в соответствии с P_k, определяется Y – скрытый класс;

Модель попарной классификации с латентными классами

- Y - непосредственно ненаблюдаемая (латентная) *переменная* (номер класса), $Y \in \{1,...,K\}$;
- *Априорные вероятности* $P_k = P(Y = k)$, $\sum_{k=1}^{K} P_k = 1$;
- *Условные распределения* $p(x \mid Y = k) = p_k(x)$, $k = 1,...,K$.
- Для объекта, в соответствии с P_k, определяется Y – скрытый класс;
- В соответствии с $p_k(x)$ определяется значение x;

Модель попарной классификации с латентными классами

- Y - непосредственно ненаблюдаемая (латентная) *переменная* (номер класса), $Y \in \{1,...,K\}$;
- *Априорные вероятности* $P_k = P(Y = k)$, $\sum_{k=1}^{K} P_k = 1$;
- *Условные распределения* $p(x \mid Y = k) = p_k(x)$, $k = 1,...,K$.
- Для объекта, в соответствии с P_k, определяется Y – скрытый класс;
- В соответствии с $p_k(x)$ определяется значение x;
- Объекты независимы.

Пусть a, b - произвольная пара различных объектов, рассмотрим

$$Z = \begin{cases} 1, Y(a) \neq Y(b) \\ 0, \ иначе \end{cases}.$$

Пусть a, b - произвольная пара различных объектов, рассмотрим

$$Z = \begin{cases} 1, Y(a) \neq Y(b) \\ 0, \ иначе \end{cases}.$$

Предположим, каждый алгоритм μ_m рандомизирован, т.е. зависит от случайного вектора Ω_m, принадлежащего некоторому множеству $\mathbf{\Omega}_m$ (параметров).

Пусть a, b - произвольная пара различных объектов, рассмотрим

$$Z = \begin{cases} 1, Y(a) \neq Y(b) \\ 0, \ иначе \end{cases}.$$

Предположим, каждый алгоритм μ_m рандомизирован, т.е. зависит от случайного вектора Ω_m, принадлежащего некоторому множеству $\mathbf{\Omega}_m$ (параметров).

Предположим, что для a, b выполняется:

$$P[h_m(\Omega_m) = 1 \mid Z = 1] = P[h_m(\Omega_m) = 0 \mid Z = 0] = q_m$$

(q_m - условная вероятность правильного решения).

Пусть a, b - произвольная пара различных объектов, рассмотрим

$$Z = \begin{cases} 1, & Y(a) \neq Y(b) \\ 0, & \textit{иначе} \end{cases}.$$

Предположим, каждый алгоритм μ_m рандомизирован, т.е. зависит от случайного вектора Ω_m, принадлежащего некоторому множеству $\mathbf{\Omega}_m$ (параметров).

Предположим, что для a, b выполняется:

$$P[h_m(\Omega_m) = 1 \mid Z = 1] = P[h_m(\Omega_m) = 0 \mid Z = 0] = q_m$$

(q_m - условная вероятность правильного решения).

Будем полагать, что $q_m > 0.5$ (условие «слабой обученности»).

Пусть алгоритм μ_m проработал L_m раз при условии выбора случайных, независимых и одинаково распределенных параметров $\Omega_{1,m},...,\Omega_{L_m,m}$; получим набор случайных решений $h_m(\Omega_{1,m}),...,h_m(\Omega_{L_m,m}),\ \ m=1,...,M$.

Пусть алгоритм μ_m проработал L_m раз при условии выбора случайных, независимых и одинаково распределенных параметров $\Omega_{1,m},...,\Omega_{L_m,m}$; получим набор случайных решений $h_m(\Omega_{1,m}),...,h_m(\Omega_{L_m,m}),\ \ m=1,...,M$.

Предположим, решения условно независимы:

$$P[h_{m_1}(\Omega_{i_1,m_1})=h_{r_1},...,h_{m_j}(\Omega_{i_j,m_j})=h_{r_j}\mid Z=z]=$$

$$P[(h_{m_1}(\Omega_{i_1,m_1})=h_{r_1}\mid Z=z]\cdot...\cdot P[h_{m_j}(\Omega_{i_j,m_j})=h_{r_j}\mid Z=z].$$

Пусть алгоритм μ_m проработал L_m раз при условии выбора случайных, независимых и одинаково распределенных параметров $\Omega_{1,m},...,\Omega_{L_m,m}$; получим набор случайных решений $h_m(\Omega_{1,m}),...,h_m(\Omega_{L_m,m}),\ \ m=1,...,M$.

Предположим, решения условно независимы:

$$P[h_{m_1}(\Omega_{i_1,m_1})=h_{r_1},...,h_{m_j}(\Omega_{i_j,m_j})=h_{r_j}\mid Z=z]=$$

$$P[(h_{m_1}(\Omega_{i_1,m_1})=h_{r_1}\mid Z=z]\cdot...\cdot P[h_{m_j}(\Omega_{i_j,m_j})=h_{r_j}\mid Z=z].$$

Обозначим

$$\overline{H}=\sum_{m=1}^{M}\alpha_m\frac{1}{L_m}\sum_{l=1}^{L_m}h_m(\Omega_{l,m})$$

- коллективная оценка с учетом весов.

Рассмотрим маргинальную функцию («отступ») кластерного ансамбля для объектов a, b:

$mg = \{$взвешенное число голосов за Z – взвешенное число голосов против $Z\}$,

где $Z \in \{0,1\}$ (a, b в «одном кластере», «разных кластерах»).

Рассмотрим маргинальную функцию («отступ») кластерного ансамбля для объектов a, b:

$mg = \{$взвешенное число голосов за Z – взвешенное число голосов против $Z\}$,
где $Z \in \{0,1\}$ (a, b в «одном кластере», «разных кластерах»).

Вероятность ошибки ансамбля для объектов a, b:

$$P_{err} = P_{Z,\Omega_{1,1},\ldots,\Omega_{L_M,M}}[mg(\overline{H}, Z) < 0].$$

Рассмотрим маргинальную функцию («отступ») кластерного ансамбля для объектов a, b:

$mg = \{$взвешенное число голосов за Z – взвешенное число голосов против $Z\}$,
где $Z \in \{0,1\}$ (a, b в «одном кластере», «разных кластерах»).

Вероятность ошибки ансамбля для объектов a, b:

$$P_{err} = P_{Z,\Omega_{1,1},\ldots,\Omega_{L_M,M}}[mg(\overline{H}, Z) < 0].$$

Пусть выполняются сформулированные выше предположения модели. Тогда

$$P_{err} < \frac{Var[mg(\overline{H}, Z)]}{(E[mg(\overline{H}, Z)])^2} \qquad (*)$$

при условии $E\, mg(\overline{H}, Z) > 0$, где

$$E[mg(\overline{H}, Z)] = 2\sum_{m=1}^{M} \alpha_m q_m - 1, \quad Var[mg(\overline{H}, Z)] = 4\sum_{m=1}^{M} \frac{\alpha_m^2}{L_m} q_m (1 - q_m).$$

Необходимо минимизировать верхнюю границу $(*)$.

Необходимо минимизировать верхнюю границу $(*)$.

Модификация критерия: будем искать

$$\alpha^* = \arg\min_{\alpha_1 \ldots \alpha_M} Var[mg(\bar{H}, Z)], \text{ s.t. } \alpha_1 \geq 0, \ldots, \alpha_M \geq 0, \sum_m \alpha_m = 1, .$$

Необходимо минимизировать верхнюю границу (*).

Модификация критерия: будем искать

$$\alpha^* = \arg\min_{\alpha_1 \ldots \alpha_M} Var[mg(\bar{H}, Z)], \text{ s.t. } \alpha_1 \geq 0, \ldots, \alpha_M \geq 0, \sum_m \alpha_m = 1, .$$

Решение методом множителей Лагранжа:

$$\alpha_m^* = \frac{\frac{L_m}{q_m(1-q_m)}}{\sum_m \frac{L_m}{q_m(1-q_m)}}, m = 1, \ldots, M.$$

Необходимо минимизировать верхнюю границу (*).

Модификация критерия: будем искать

$$\alpha^* = \arg\min_{\alpha_1 \ldots \alpha_M} Var[mg(\bar{H}, Z)], \text{ s.t. } \alpha_1 \geq 0, \ldots, \alpha_M \geq 0, \sum_m \alpha_m = 1, .$$

Решение методом множителей Лагранжа:

$$\alpha_m^* = \frac{\frac{L_m}{q_m(1-q_m)}}{\sum_m \frac{L_m}{q_m(1-q_m)}}, m = 1, \ldots, M.$$

Байесовские оценки вероятности отнесения пары объектов к разным кластерам: $\tilde{q}_m^B = \max(\tilde{P}_m^B, 1 - \tilde{P}_m^B)$, где $\tilde{P}_m^B = \frac{\sum_l \bar{h}_{l,m} + 1}{L_m + 2}$.

Algorithm Pairwise Weighted Ensemble Clustering (PWEC)

Input

$\mathbf{x}_N = \{x_1, ..., x_N\}$ таблица данных в $\mathbf{R}^d$,

K: число кластеров

$L_1, ..., L_M$: число запусков алгоритмов $\mu_1, ..., \mu_M$;

Output

разбиение $\mathbf{x}_N$ на K кластеров;

Procedure

1. **for m:=1 to M do**
2. Генерировать L_m вариантов группировки алгоритмом μ_m для случайно выбранных условий работы;

end;

3. **Для каждой пары различных объектов** a, b **из** $\mathbf{x_N}$ **do**
4. Вычислить оценки $\overline{h}_{l,m}$, $l = 1, ..., L_m$, $m = 1, ..., M$;
5. Найти оптимальные веса $\alpha_1^*, ..., \alpha_M^*$;

end;

6. Вычислить матрицу попарных различий $\mathbf{H}$, используя найденные веса;
7. Применить агломеративный алгоритм построения иерархического дерева группировки, используя матрицу $\mathbf{H}$ на входе.

end.

Экспериментальное исследование

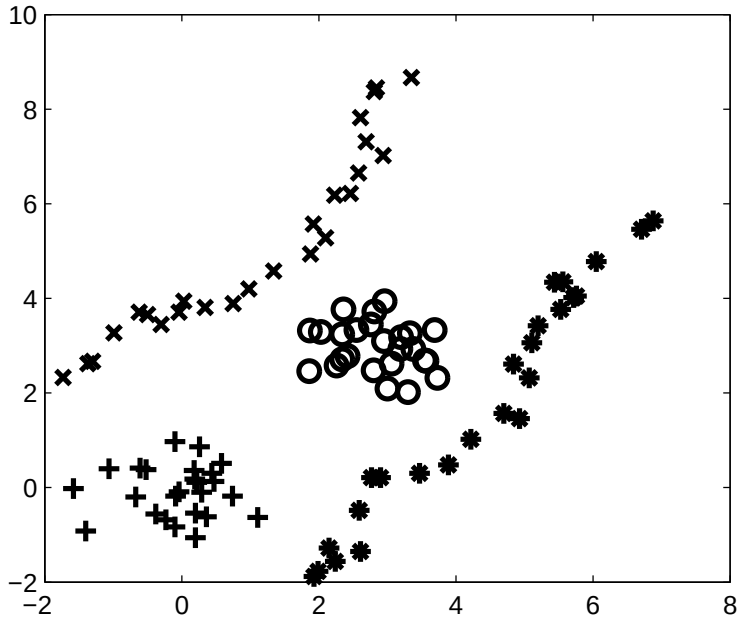
В ансамбль были включены 2 алгоритма:

- агломеративный алгоритм построения иерархической группировки (дендрограммы), использующий расстояние по «ближайшему соседу» и

- алгоритм k-means

Тестовая задача

- 30-мерное пространство переменных:
- за основу взят пример:



проекция на X_1, X_2

Статистическое моделирование:

- многократное генерирование выборок, соответствующих заданному распределению для кластеров;

Статистическое моделирование:

- многократное генерирование выборок, соответствующих заданному распределению для кластеров;
- группировка с помощью алгоритма с весами;

Статистическое моделирование:

- многократное генерирование выборок, соответствующих заданному распределению для кластеров;
- группировка с помощью алгоритма с весами;
- вычисление степени согласованности полученной классификации с истинной (исправленный индекс Ранда, ARI).

Статистическое моделирование:

- многократное генерирование выборок, соответствующих заданному распределению для кластеров;
- группировка с помощью алгоритма с весами;
- вычисление степени согласованности полученной классификации с истинной (исправленный индекс Ранда, ARI).

Часть переменных - «шумовые» (с равномерным распределением)

Статистическое моделирование:

- многократное генерирование выборок, соответствующих заданному распределению для кластеров;
- группировка с помощью алгоритма с весами;
- вычисление степени согласованности полученной классификации с истинной (исправленный индекс Ранда, ARI).

Часть переменных - «шумовые» (с равномерным распределением)

Построения ансамбля - метод случайных подпространств ($d' = 3$).

Статистическое моделирование:

- многократное генерирование выборок, соответствующих заданному распределению для кластеров;
- группировка с помощью алгоритма с весами;
- вычисление степени согласованности полученной классификации с истинной (исправленный индекс Ранда, ARI).

Часть переменных - «шумовые» (с равномерным распределением)

Построения ансамбля - метод случайных подпространств ($d' = 3$).
Число элементов ансамбля $L_1 = L_2 = 50$.

Статистическое моделирование:

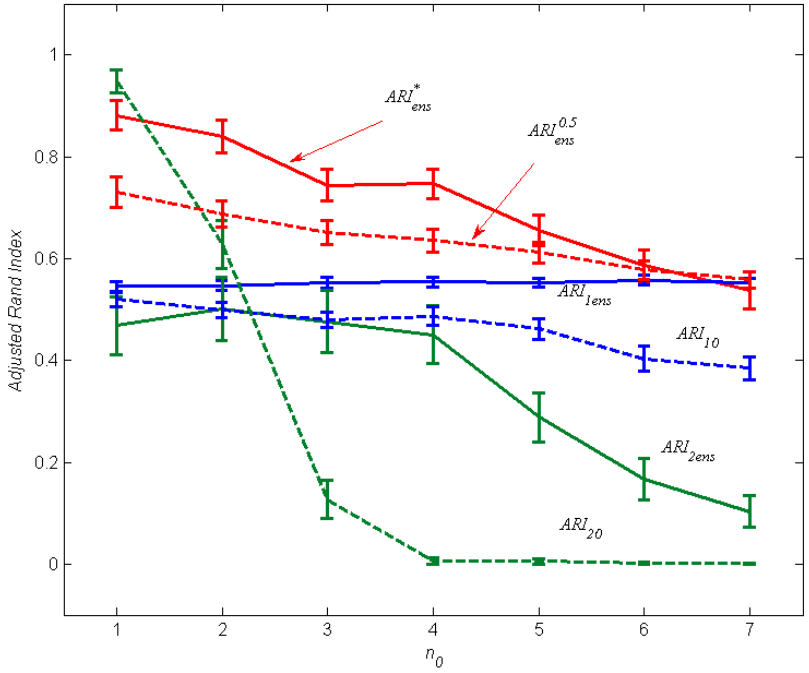
- многократное генерирование выборок, соответствующих заданному распределению для кластеров;
- группировка с помощью алгоритма с весами;
- вычисление степени согласованности полученной классификации с истинной (исправленный индекс Ранда, ARI).

Часть переменных - «шумовые» (с равномерным распределением)

Построения ансамбля - метод случайных подпространств ($d' = 3$).
Число элементов ансамбля $L_1 = L_2 = 50$.

Степень согласованности усредняется по выборкам.

Результаты моделирования (усредненный по 100 выборкам AR-индекс, в зависимости от числа шумовых переменных)



R^*_{ens} - ансамблевый алгоритм с оптимизируемыми весами;

$R^{0.5}_{ens}$- ансамблевый алгоритм с равными весами;

R_{10} - k-means;

R_{20} - агломеративный алгоритм (расстояние по ближ. соседу);

$R_{1\,ens}$ - коллектив k-means;

$R_{2\,ens}$- коллектив агломеративных алгоритмов

с удвоенным числом элементов в ансамбле

Спасибо за внимание!