Ансамбль алгоритмов кластерного анализа с весами и его применение при анализе изображений

В.В.Бериков

Институт математики им. В.Л.Соболева СО РАН

И.А.Пестунов

Институт вычислительных технологий СО РАН

G. Gonzalez

Ecole des Ponts ParisTech

Коллективный подход в кластерном анализе

+ :

- Повышается устойчивость результатов по отношению к выбору параметров работы алгоритма;
- Доказано, что при выполнении определенных условий качество группировки улучшается;
- Возможность проведения распределенных вычислений (при различном местоположении подмножеств объектов или переменных).

Два основных направления в построении композиций (ансамблей) алгоритмов

1. Нахождение консенсусного разбиения

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, $l = 1, ..., L$; $\sum_l w_l = 1$.

2. Вычисление согласованной матрицы сходства/различий (co-occurrence matrix)

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, $i, j = 1, 2, \ldots, N$, $i \neq j$.

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 иерархической 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;
- учитывается мера разнообразия вариантов: похожим вариантам разбиения приписывается меньший вес.

Выбор оптимальных весов нескольких алгоритмов 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), $Y \in \{1,...,K\}$;

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).

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

Пусть алгоритм μ_m проработал L_m раз при выборе случайных, независимых и одинаково распределенных параметров $\Omega_{1,m},\ldots,\Omega_{L_m,m}$.

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:

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

Утверждение. $P_{err} < \dfrac{Var[mg(\overline{H}, Z)]}{(E[mg(\overline{H}, Z)])^2}$,

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 ошибки.

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:

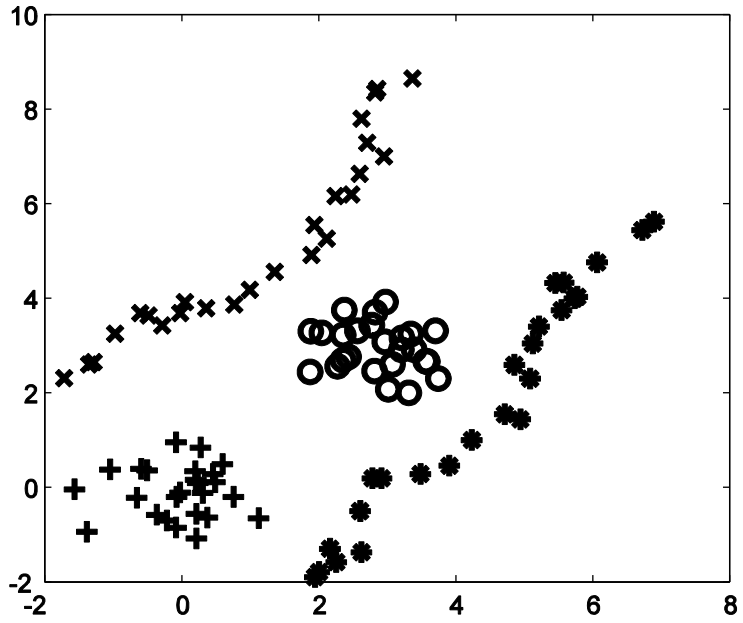
$$\alpha_m^* = \frac{\frac{L_m}{q_m(1-q_m)}}{\sum_m \frac{L_m}{q_m(1-q_m)}}, m = 1, \ldots, M.$$

Частотные оценки q_m - наблюдаемые величины!

Экспериментальное исследование алгоритма

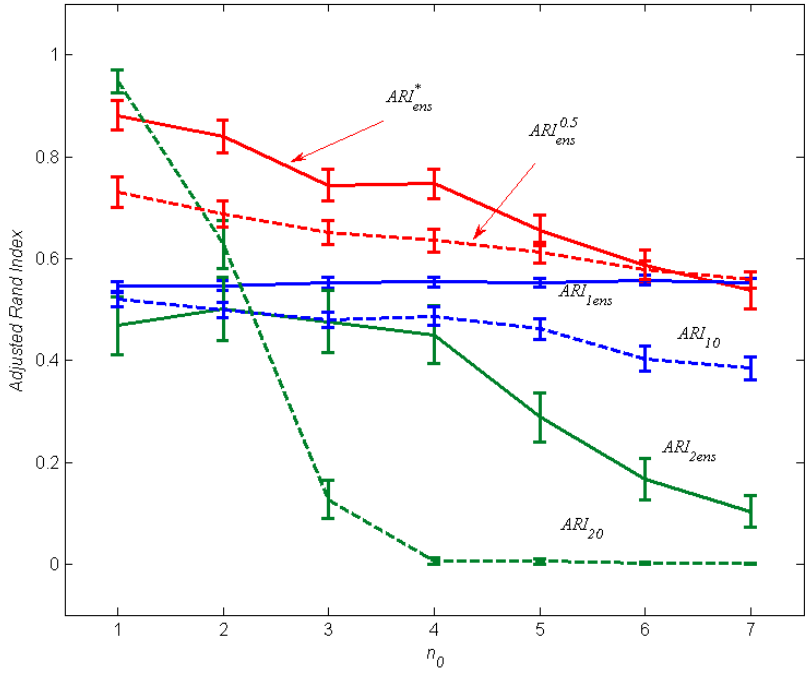
Тестовая задача

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



проекция на X_1, X_2

Результаты моделирования (усредненный по 100 выборкам индекс качества, 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;

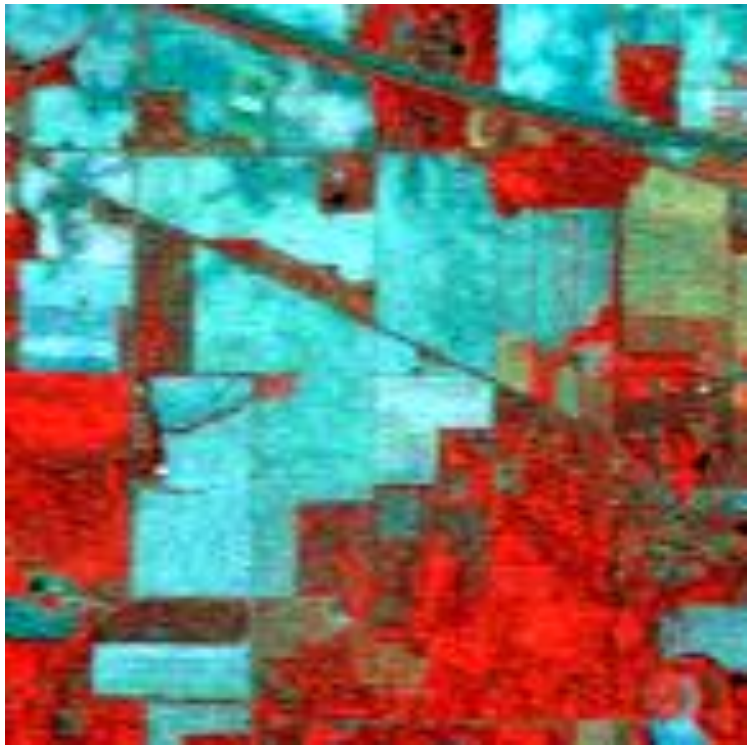
R_{20} - агломеративный алгоритм (расстояние по ближ. соседу);

$R_{1\,ens}$ - коллектив k-means;

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

с удвоенным числом элементов в ансамбле

Применение ансамблевого алгоритма для сегментации гиперспектральных изображений



изображение «Indian Pines»

особенность гиперспектральных изображений

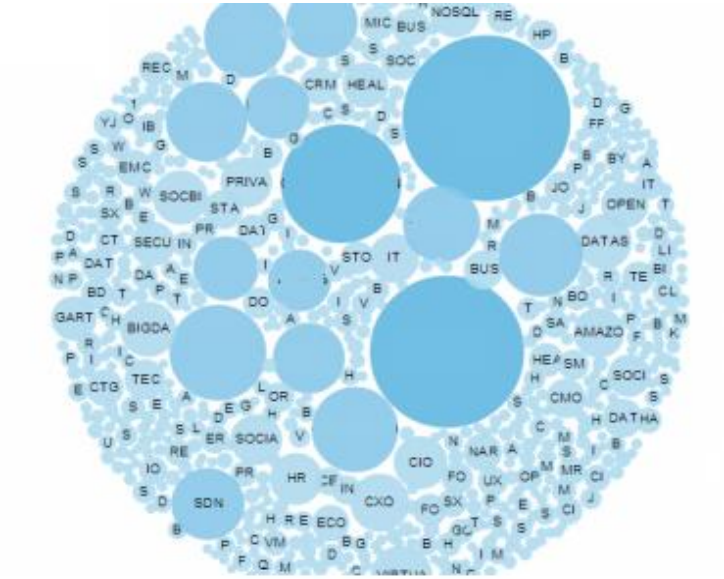
- большая размерность:

до 10^6 объектов (пикселей);

несколько сотен переменных (каналов)

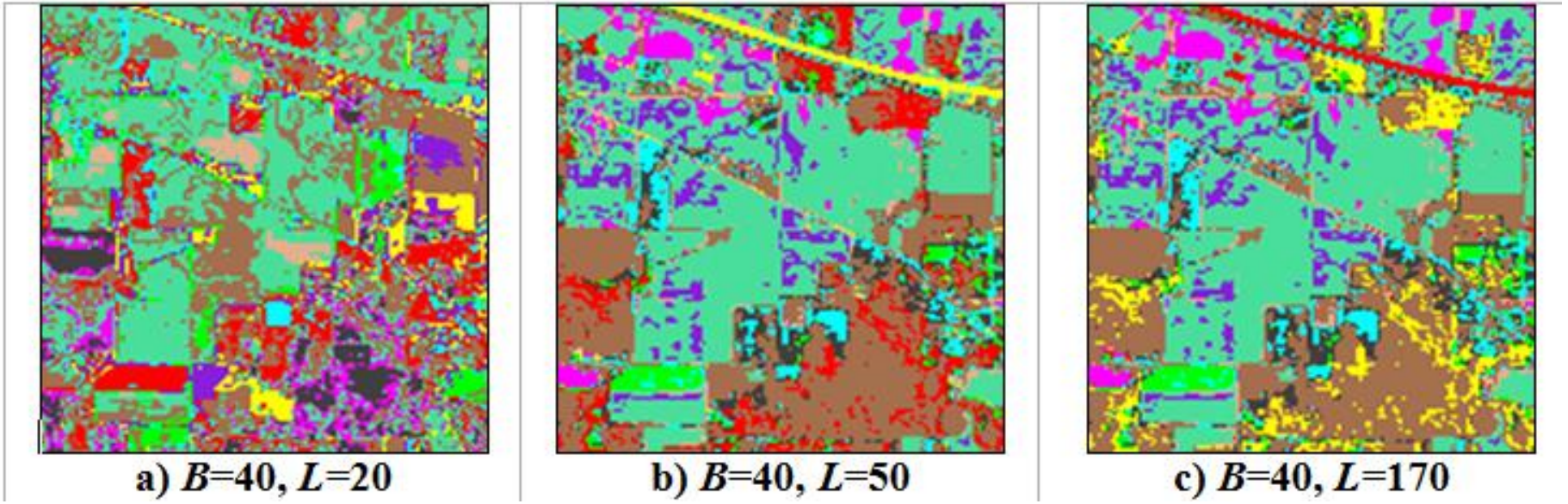
Использование алгоритма напрямую невозможно (попарные сравнения всех объектов – огромная трудоемкость!)

центроидный метод:



1. Получить варианты разбиения данных на достаточно большое число кластеров каким-либо «быстрым» алгоритмом (например, k-means);
2. Построить коллективное решение для центроидов или «прототипов» кластеров
3. Вернуться к исходным данным

Пример результатов работы алгоритма (K=10) “random subspace” ensemble (B – число каналов)



a) B=40, L=20

b) B=40, L=50

c) B=40, L=170

Направления дальнейших исследований:

- учет индексов качества группировки при вычислении весов;

- теоретическое исследование центроидного метода;

- кластеризация разнородных данных:
  многомерных разнотипных временных рядов,
  логических экспертных высказываний

Основные результаты:

- Теоретически исследована вероятностная модель кластерного ансамбля;

- Разработан алгоритм коллективной группировки с оптимизируемыми весами алгоритмов;

- Предложен алгоритм ансамблевого анализа гиперспектральных изображений

Спасибо за внимание!