Ансамблевая кластеризация: методы, эксперименты и приложения

Шестаков Андрей

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Факультет Компьютерных Наук
Департамент анализа данных и искусственного интеллекта

Содержание

Результаты применения алгоритмов клатеризации

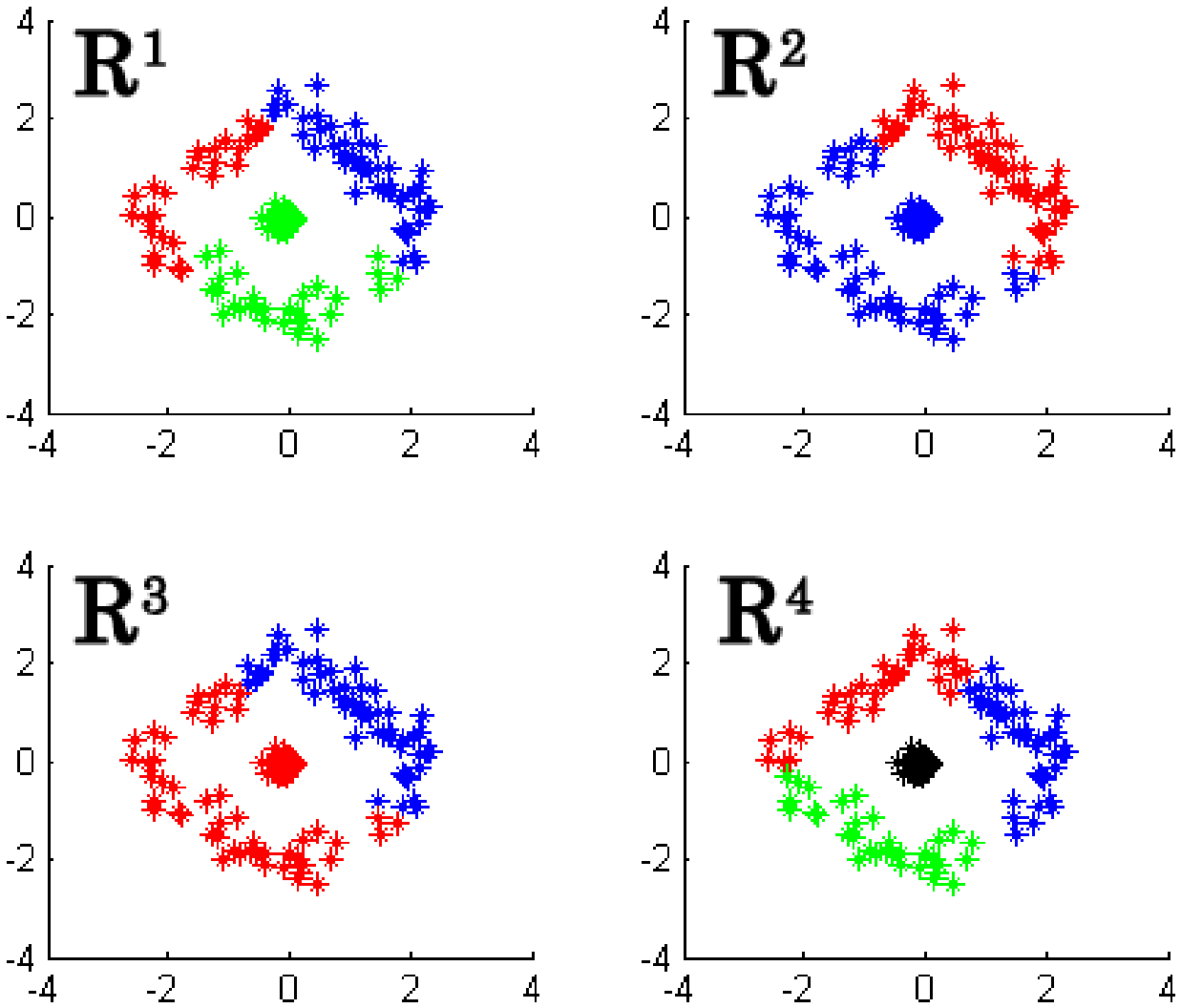


Рис. 1 : *Четыре различных разбиения*

Задача формирования согласованного разбиения

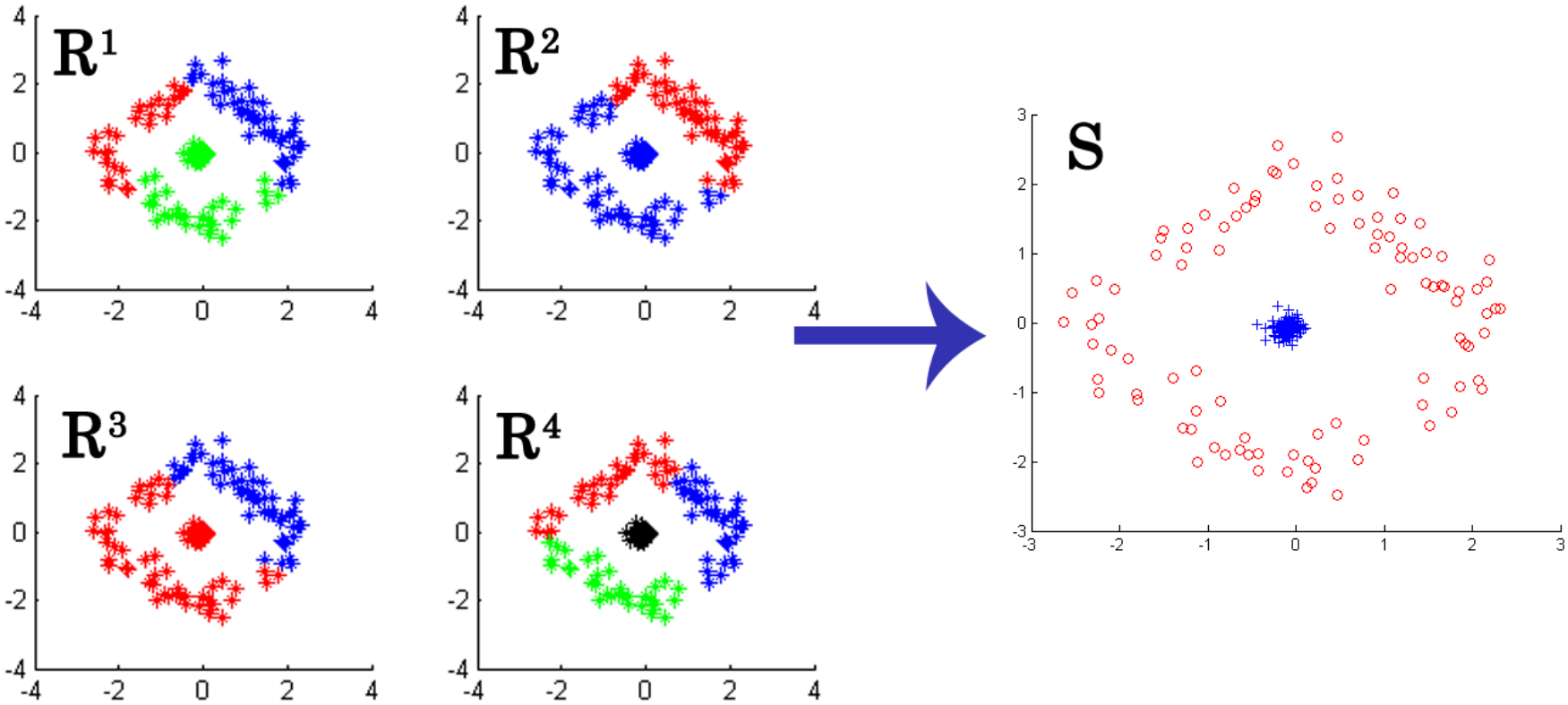


Рис. 2 : *Ансамбль разбиений (слева) и согласованное разбиение (справа)*

Задача формирования согласованного разбиения

Содержание

Голосование

Пример голосования

1. Выбираем правило голосования
2.

	R_1	R_2	R_3	R_4
y_1 :	1	1	1	1
y_2 :	1	1	2	1
y_3 :	2	2	1	2
y_4 :	2	2	2	3
y_5 :	2	3	3	2
y_6 :	3	3	3	3

$$\Rightarrow S = \begin{matrix} y_1: \\ y_2: \\ y_3: \\ y_4: \\ y_5: \\ y_6: \end{matrix} \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 3 \end{bmatrix}$$

Голосование

Проблема переобозначения меток

	R_1	R_2	R_3	R_4
y_1 :	2	1	1	1
y_2 :	2	1	2	1
y_3 :	1	2	1	2
y_4 :	1	2	2	3
y_5 :	1	3	3	2
y_6 :	3	3	3	3

$$\Rightarrow S = \begin{matrix} y_1 : \\ y_2 : \\ y_3 : \\ y_4 : \\ y_5 : \\ y_6 : \end{matrix} \begin{bmatrix} ? \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix}$$

Голосование

Проблема переобозначения меток

	R_1	R_2	R_3	R_4
y_1 :	2	1	1	1
y_2 :	2	1	2	1
y_3 :	1	2	1	2
y_4 :	1	2	2	3
y_5 :	1	3	3	2
y_6 :	3	3	3	3

$$\Rightarrow S = \begin{matrix} y_1: \\ y_2: \\ y_3: \\ y_4: \\ y_5: \\ y_6: \end{matrix} \begin{bmatrix} ? \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix}$$

Решение

1. Выбираем ключевое разбиение
2. Венгерский алгоритм (Hungarian algorithm)

Ссылки

- Kuhn, Yaw 1955
- Weingessel, Dimitriadou and Hornik, 2001
- Sevillano, Cobo, Alías and Socoró, 2006
- Ayad, 2010
- ...

Попарное сходство

Консенсусная матрица

	R_1	R_2	R_3	R_4
y_1 :	2	1	1	1
y_2 :	2	1	2	1
y_3 :	1	2	1	2
y_4 :	1	2	2	3
y_5 :	1	3	3	2
y_6 :	3	3	3	3

$\Leftrightarrow A =$

	y_1	y_2	y_3	y_4	y_5	y_6
y_1	4	3	1	0	0	0
y_2		4	0	1	0	0
y_3			4	2	0	0
y_4				4	1	1
y_5					4	2
y_6						4

Модификации/Вариации

Iam-on and Garrett, 2010

- Sim-Rank-based similarity

$$s(y_i, y_j) = \frac{\beta}{|N_{y_i}||N_{y_j}|} \sum_{a \in N_{y_i}} \sum_{b \in N_{y_i}} s(a, b)$$

- Connected-triple-based similarity

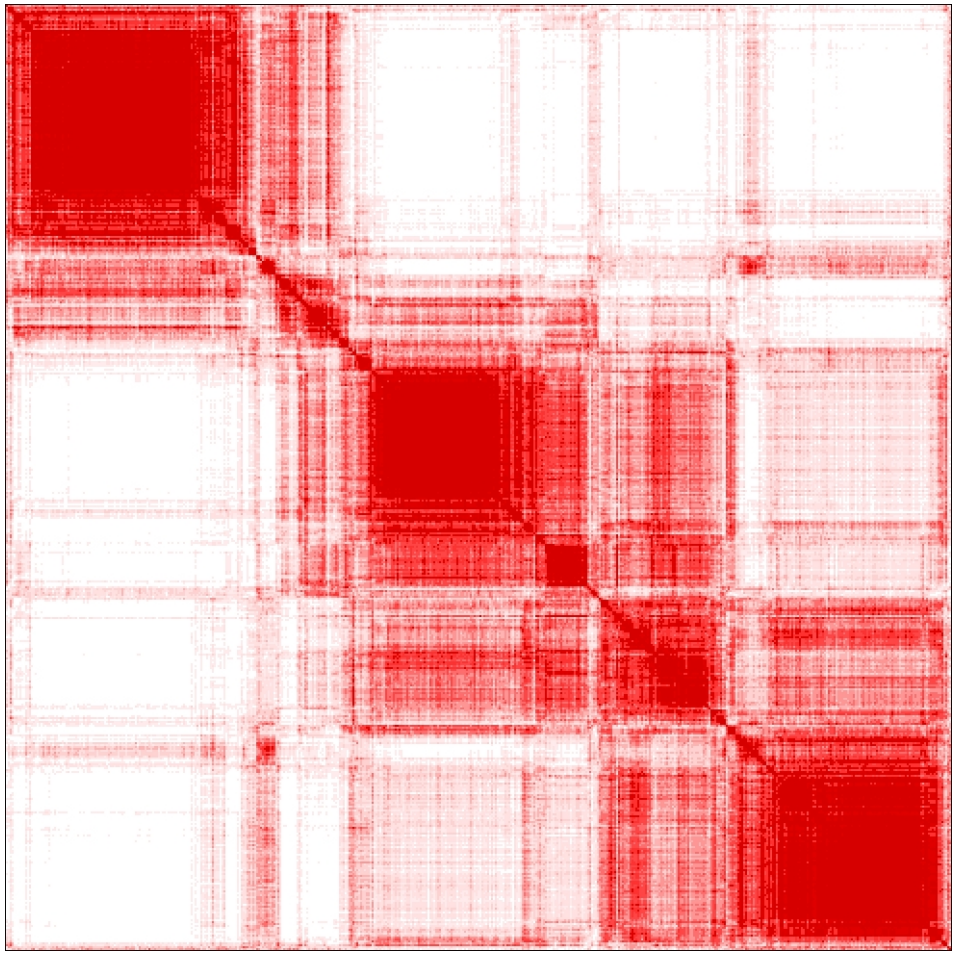


Рис. 3 : *Пример консенсусной матрицы*

- Fred and Jain, 2005
- Guènoche 2011
- ...

Графовые методы

1. Представить профиль разбиений виде графовой структуры (гиперграф, двудольный граф)
2. Найти минимальные резрезы

	$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$
x_1	1	2	1	1
x_2	1	2	1	2
x_3	1	2	2	?
x_4	2	3	2	1
x_5	2	3	3	2
x_6	3	1	3	?
x_7	3	1	3	?

$\Leftrightarrow$

	$\mathbf{H}^{(1)}$			$\mathbf{H}^{(2)}$			$\mathbf{H}^{(3)}$			$\mathbf{H}^{(4)}$	
	$\mathbf{h}_1$	$\mathbf{h}_2$	$\mathbf{h}_3$	$\mathbf{h}_4$	$\mathbf{h}_5$	$\mathbf{h}_6$	$\mathbf{h}_7$	$\mathbf{h}_8$	$\mathbf{h}_9$	$\mathbf{h}_{10}$	$\mathbf{h}_{11}$
v_1	1	0	0	0	1	0	1	0	0	1	0
v_2	1	0	0	0	1	0	1	0	0	0	1
v_3	1	0	0	0	1	0	0	1	0	0	0
v_4	0	1	0	0	0	1	0	1	0	1	0
v_5	0	1	0	0	0	1	0	0	1	0	1
v_6	0	0	1	1	0	0	0	0	1	0	0
v_7	0	0	1	1	0	0	0	0	1	0	0

Рис. 4 : *Гипер-граф, полученный из разбиения*

- Strehl and Ghosh, 2002
- Fern and Brodley, 2004
- Ng *et al*, 2004

Вероятностные методы

- Mixture models - Tophy, *et al* 2005
- Nonparametric Bayesian Co-clustering Ensembles - Wang, *et al* 2009
- Probabilistic consensus clustering - Laurenco, *et al* 2014

Содержание

Критерий наименьших квадратов

Разбиение и матрица инцидентности

$$S = \{S_1, S_2, S_3\} = \begin{matrix} y_1: \\ y_2: \\ y_3: \\ y_4: \\ y_5: \\ y_6: \end{matrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \\ 2 \\ 2 \end{bmatrix} \Leftrightarrow Z =$$

	S_1	S_2	S_3
$y_1:$	1	0	0
$y_2:$	0	1	0
$y_3:$	0	0	1
$y_4:$	1	0	0
$y_5:$	0	1	0
$y_6:$	0	1	0

Проекция на $L(Z)$

$$P_z = Z(Z^T Z)^{-1} Z^T = (p_{ij})$$

$$p_{ij} = \begin{cases} \frac{1}{|S_k|}, & \text{если } \{y_i, y_j\} \in S_k; \\ 0, & \text{иначе.} \end{cases}$$

Два подхода к задаче (Mirkin, Muchnik (1981), Mirkin (2012))

Даны разбиения $R^1, R^2, ..., R^T$, найти такое согласованное разбиение S что:

- Ensemble consensus: S хорошо восстанавливает R^t, $t = 1, 2, \dots T$
- Combined consensus: R^t, $t = 1, 2, \dots T$ хорошо восстанавливают S

Критерий наименьших квадратов

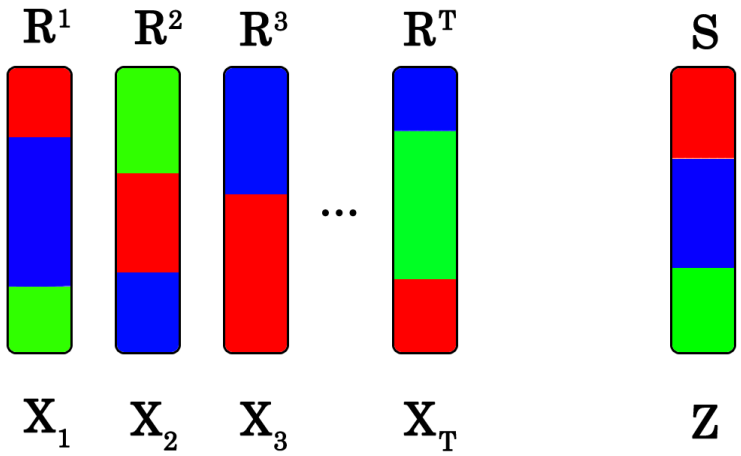


Рис. 5 : *Разбиения* $S, R^1, \ldots R^T$ *и соответствующие матрицы* $Z, X_1, \ldots, X_T$

Ensemble consensus

$$S \Rightarrow \{R^1, R^2, \ldots, R^T\}$$
$$\Updownarrow$$
$$E^2 = \sum_{t=1}^{T} \|X_t - P_Z X_t\|^2$$

Combined consensus

$$\{R^1, R^2, \ldots, R^T\} \Rightarrow S$$
$$\Updownarrow$$
$$E^2 = \sum_{u=1}^{T} \|Z - P_u Z\|^2$$

..и после некоторых выводов

Ensemble consensus clustering

$$g(S) = \sum_{k=1}^{K} \sum_{i,j \in S_k} a_{ij}/|S_k|$$

где $A = (a_{ij})$ — консенсусная матрица разбиений $R = \{R^1, \ldots, R^T\}$.

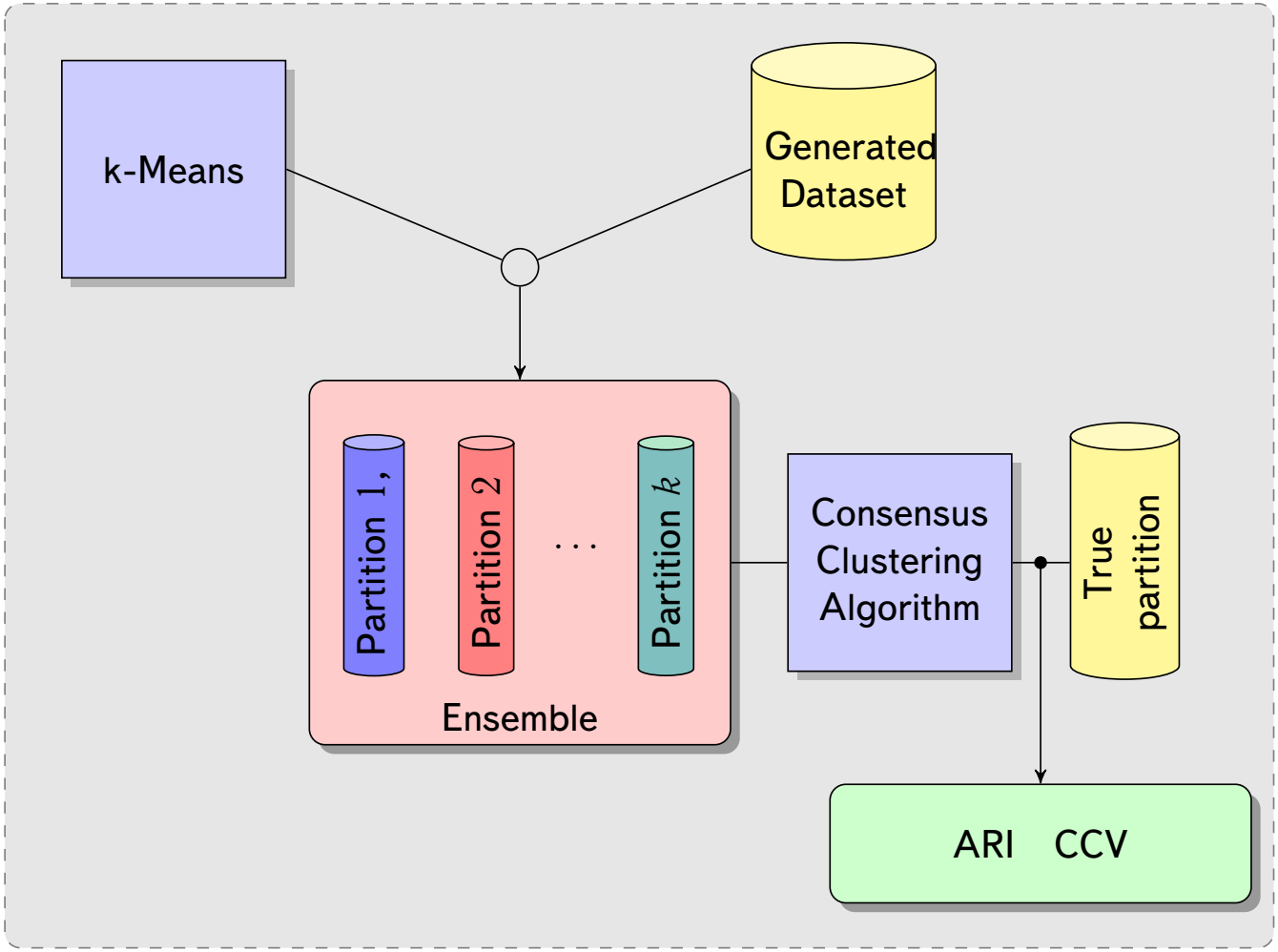
Combined consensus clustering

$$f(S) = \sum_{k=1}^{K} \sum_{i,j \in S_k} (p_{ij} - T/N)$$

где $P = (p_{ij})$ — сумарный проектор на разбиения $R = \{R^1, \ldots, R^T\}$, а N – кол-во объектов.

Содержание

Сгенерированные данные



Сгенерированные данные

- Сферические гауссианы: 1000 наблюдений, 12 признаков, 9 кластеров
- Размеры кластеров
  - Равномощные
  - Равномерно случайные
  - Заданные доли 30/20/10/10/6/6/6/6/6

Сравнение методов

- Bayesian Cluster Ensemble (Wang - 2009)
- Voting Scheme (Dimitriadou, Weingessel and Hornik - 2002)
- cVote (Ayad - 2010)
- Fusion-Transfer (Guenoche - 2011)
- Borda Consensus (Sevillano, Carrie and Pujol - 2008)
- Meta-CLustering Algorithm (Strehl and Ghosh - 2002)
- Hyper Graph Partitioning Algorithm (Strehl and Ghosh - 2002)
- Cluster-based Similarity Partitioning Algorithm (Strehl and Ghosh - 2002)

Равномощные кластеры

Similar results for other cluster models

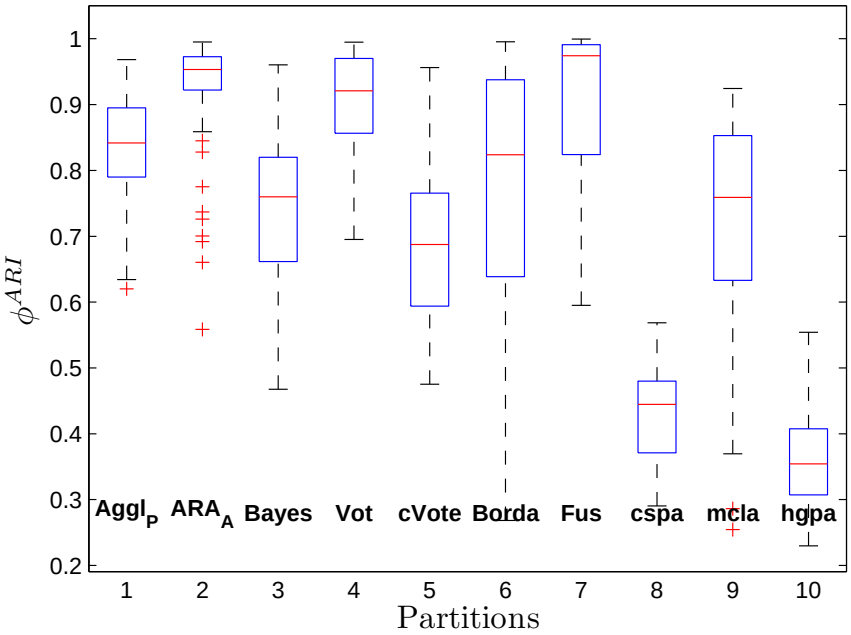


Рис. 6 : *Adjusted Rand Index boxplot*

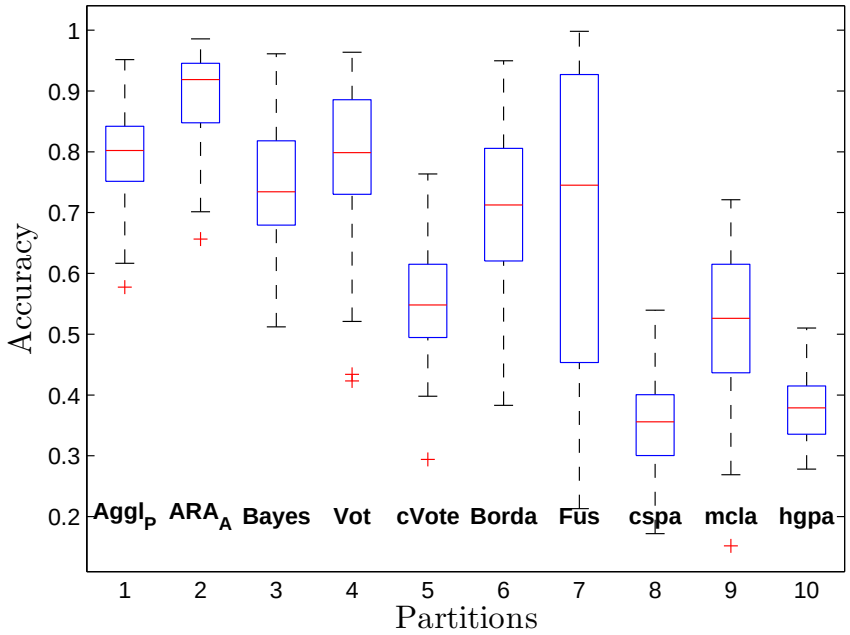
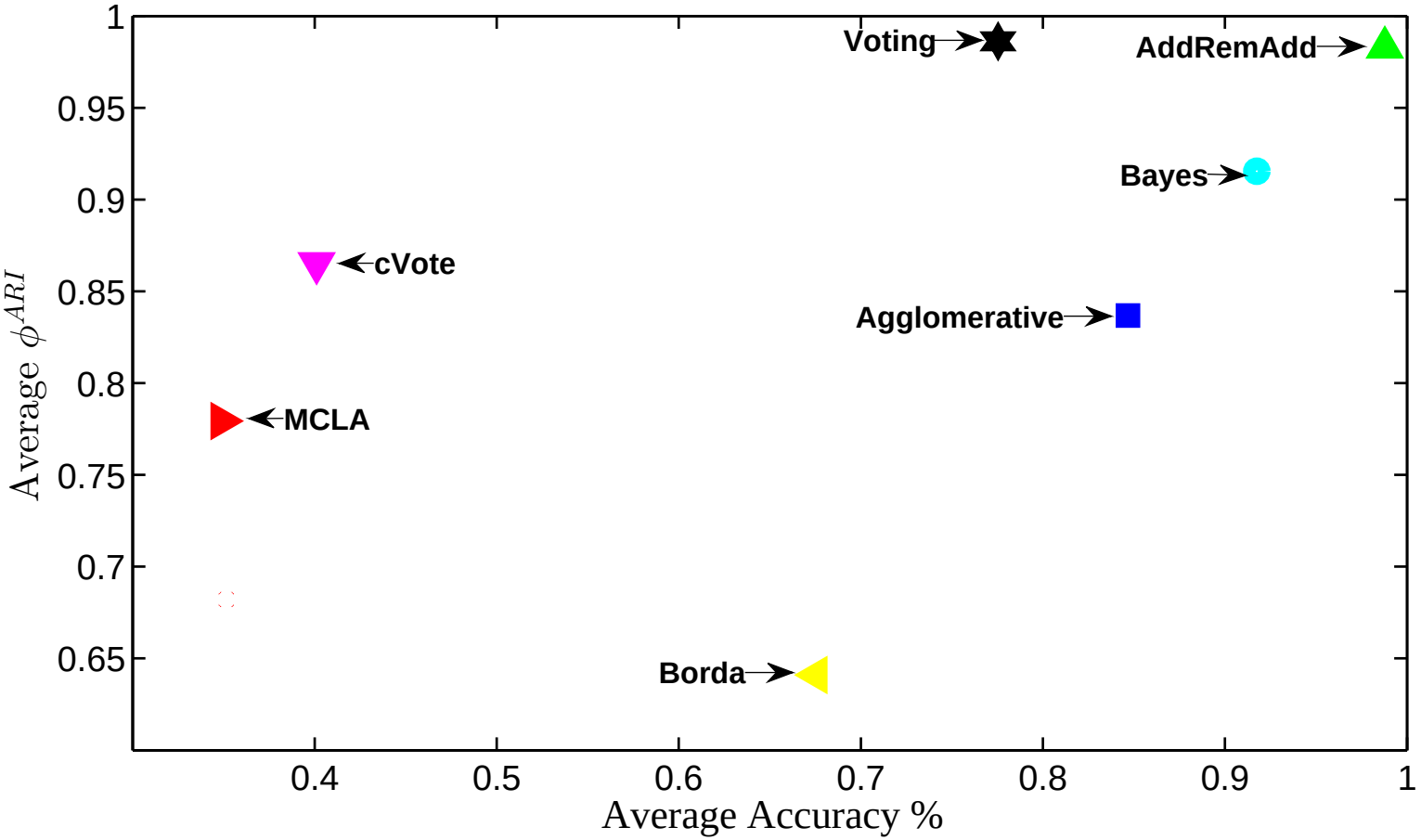


Рис. 7 : *Accuracy boxplot*

Равномощные кластеры II

ϕ^{ARI}	Aggl	ARA	Bayes	Vot	cVote	Borda	mcla
Std.	0.0799	**0.0319**	0.0726	**0.0153**	0.0451	0.0655	0.0573

Accuracy	Aggl	ARA	Bayes	Vot	cVote	Borda	mcla
Std.	0.0791	**0.0293**	0.0762	0.0795	0.1085	**0.0452**	0.0728



Заданные пропорции

Среднее значение ARI & Accuracy по 50 запускам

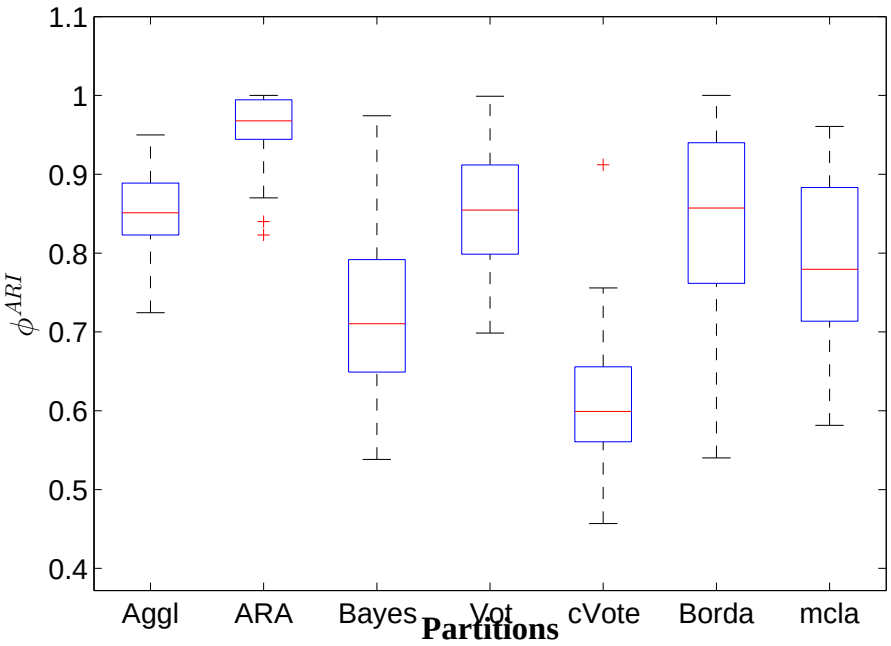


Рис. 8 : *ARI boxplot*

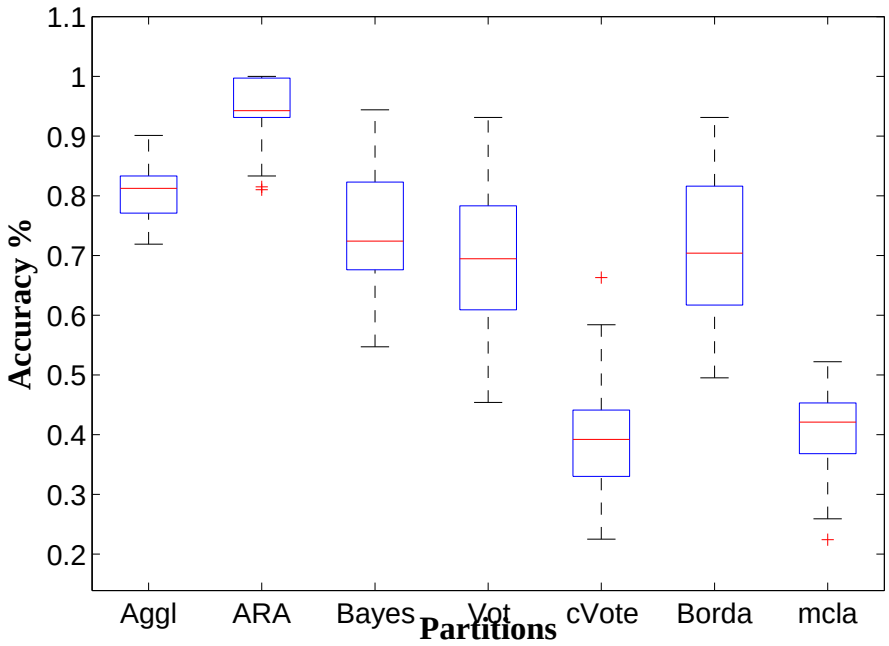
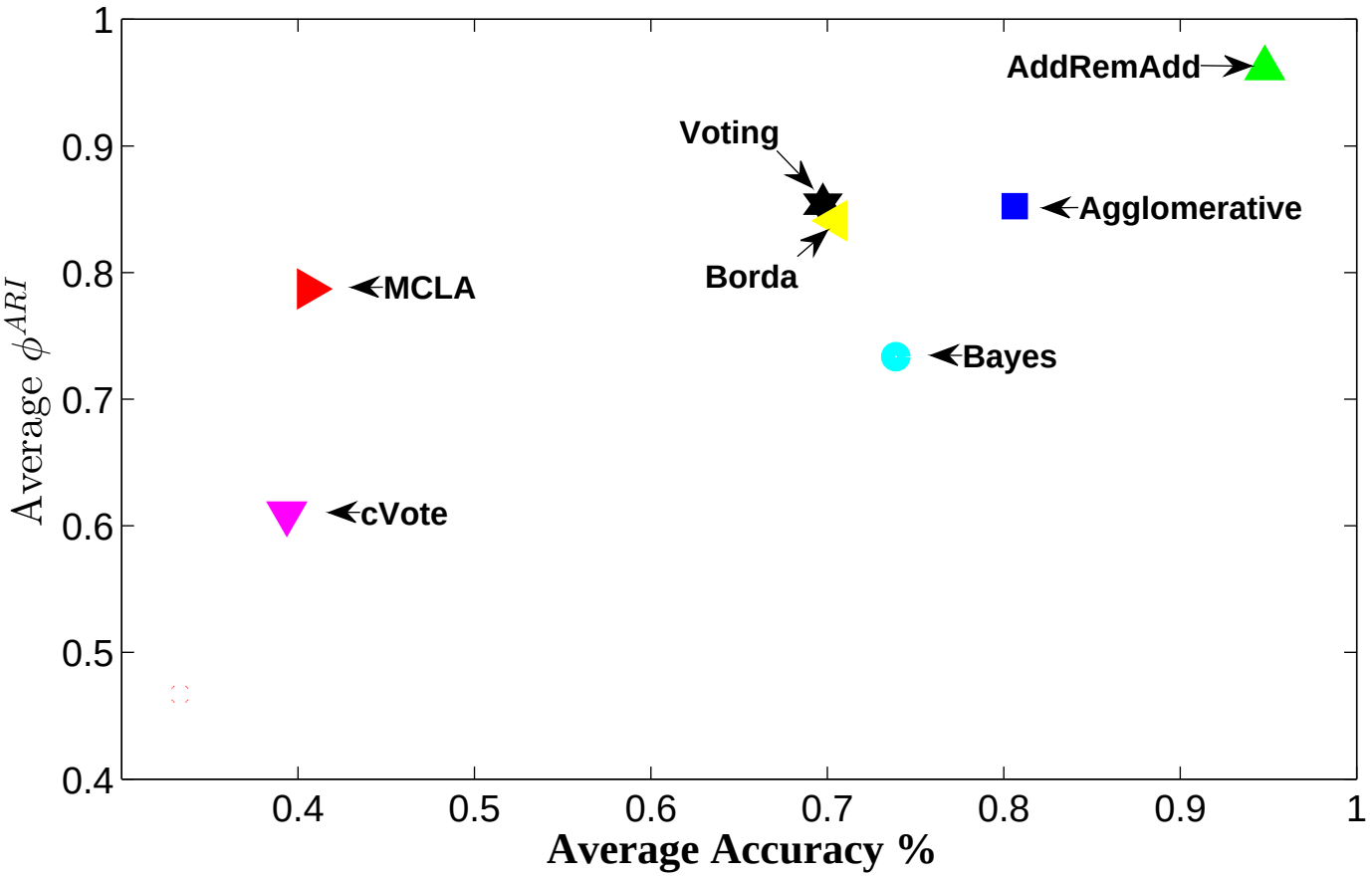


Рис. 9 : *Accuracy boxplot*

Заданные пропорции

ϕ^{ARI}	Aggl	ARA	Bayes	Vot	cVote	Borda	mcla
Std.	**0.0486**	**0.0410**	0.1046	0.0813	0.0784	0.1117	0.0992

Accuracy	Aggl	ARA	Bayes	Vot	cVote	Borda	mcla
Std.	**0.0400**	**0.0524**	0.0963	0.1189	0.0843	0.1177	0.0671



Равномерно случайные размеры кластеров

Среднее значение ARI & Accuracy за 50 запусков

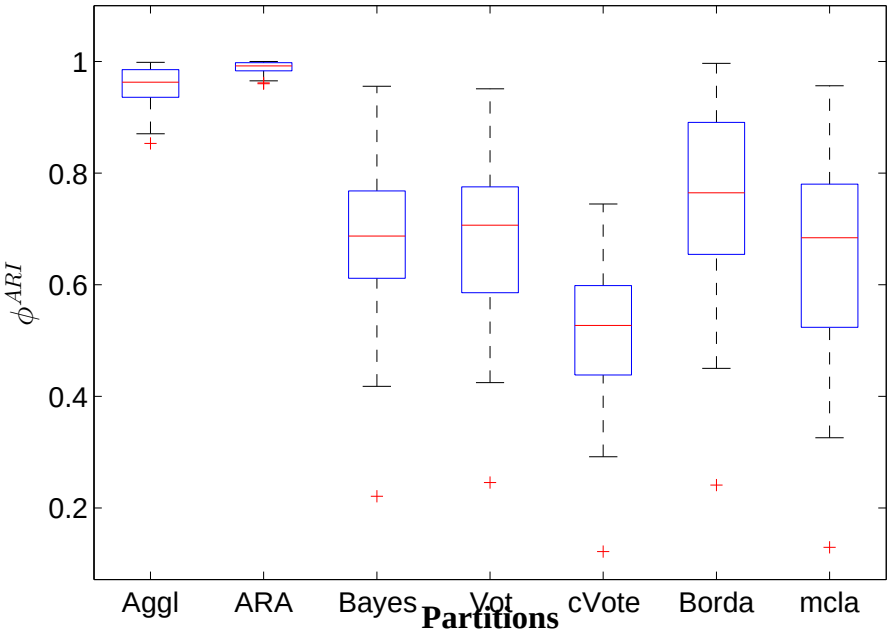


Рис. 10 : *Adjusted Rand Index boxplot*

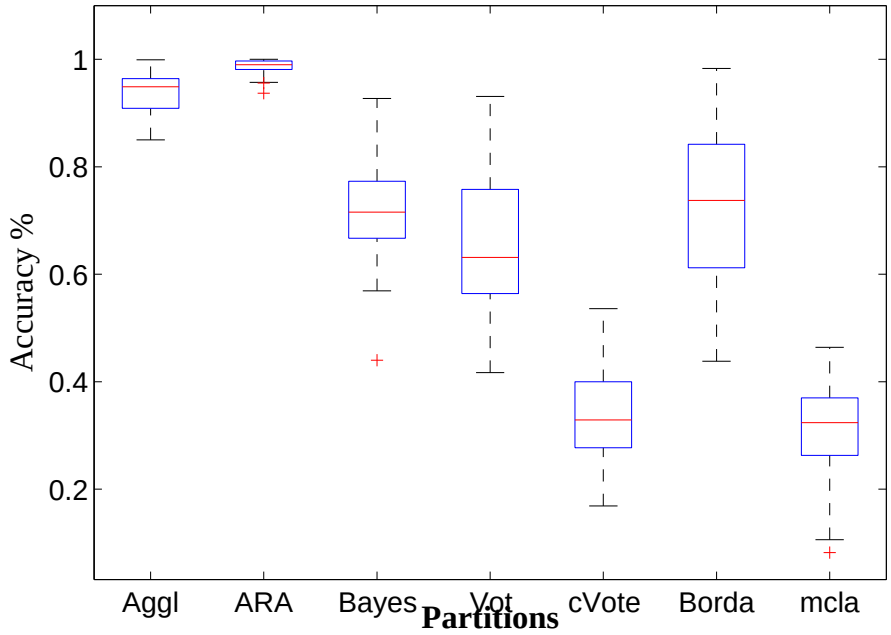


Рис. 11 : *Accuracy boxplot*

Содержание

Консенсусная кластеризация и community detection

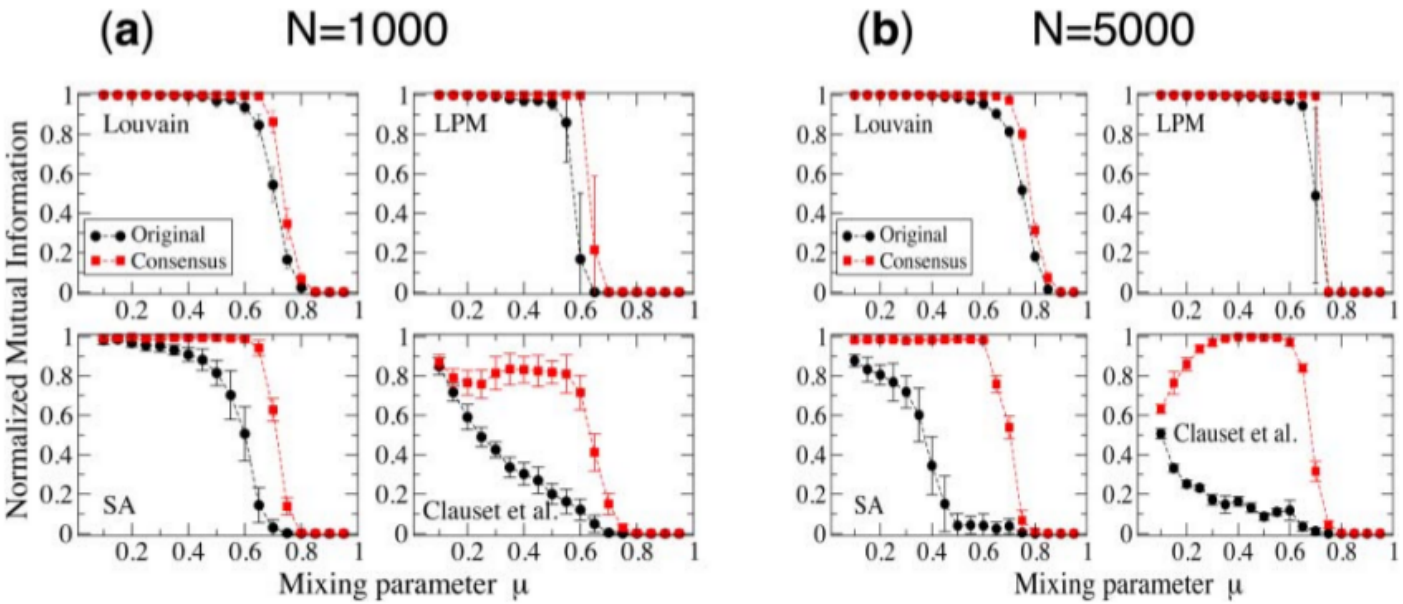


Рис. 12 : *Consensus clustering in complex networks. Fortunato and Lancichinetti, 2012*

Спасибо за внимание!