**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра МОЭВМ**

**отчет**

**по лабораторной работе №7**

**по дисциплине «Методы статистической обработки данных»**

**Тема: Кластерный анализ. Метод поиска сгущений.**

|  |  |  |
| --- | --- | --- |
| Студент гр. 5381 |  | Лянгузов А. А. |
| Преподаватель |  | Середа В. И. |

Санкт-Петербург

2019

**Цель работы.**

Освоение основных понятий и некоторых методов кластерного анализа.

**Основные теоретические положения.**

*Кластерный анализ* (или кластеризация) – задача распределения однородных объектов из данного набора по группам (кластерам) таким образом, что объекты, принадлежащие одной группе, больше похожи друг на друга, чем на объекты из других групп. Процесс кластеризации подразумевает формирование этих групп. Это общая нечеткая формулировка задачи, на практике существует множество критериев «похожести» объектов – требований, предъявляемых кластерам, поэтому существует большое количество различных алгоритмов кластеризации, а также их модификаций.

Существующие методы можно разделить следующим образом:

1. По степени принадлежности объектов кластерам:
   * Строгая кластеризация – каждый объект может принадлежать только одному кластеру.
   * Нестрогая кластеризация – каждый объект может принадлежать нескольким кластерам в разной степени.
2. По способу формирования кластеров (по критериям кластеров):
   * Иерархические алгоритмы (по связанности объектов)
     1. Аггломеративные (восходящие, объекты объединяются в кластеры)
     2. Дивизивные (нисходящие, кластеры дробятся на более мелкие)
   * Алгоритмы, основанные на поиске/использовании центров кластеров (k-means)
   * Кластеризация по распределениям
   * Кластеризация по плотностям

В данной лабораторной работе использовалось два *критерия качества* кластерных разбиений:

* + 1. Средний кластерный радиус :

Для данного показателя вычисляются для каждого кластера расстояния от центров до точек, им принадлежащих, находятся средние радиусы для каждого кластера, а затем вычисляется среднее значение из средних.

* + 1. Среднее внутрикластерное расстояние :

Положим, в кластере содержится элементов, тогда для каждого кластера вычисляются расстояния между его элементами и высчитывается среднее, а затем для всех кластеров находится среднее из средних.

*Метод поиска сгущений (ForEl).*

В данном методе каждый кластер определяется своим центром (как и в к-средних) и радиусом (одинаковым для всех кластеров), которые вместе образуют круг (шар). Эти области покрывают все множество исходных точек. Данному кластеру принадлежат точки, лежащие внутри его круга.

Очевидно, что ситуация, когда области нескольких кластеров пересекаются, может возникнуть довольно часто. Необходимо обуславливаться, каким образом решать подобные конфликты. В данной работе исследовалась строгая кластеризация и предпочтение отдавалось кластеру с наибольшим количеством элементов, при равенстве элементов выбирался первый встретившийся кластер из соперничающих.

Входные данные:

* исходное множество, подлежащее кластеризации;
* R – радиус.

Выбор кластерного радиуса является нетривиальной задачей, поскольку зависит от вида кластеризуемого множества.

Также алгоритму входе поиска кластеров (итеративно) необходимо выбирать стартовые точки для инициализации центров. Конечный результат может сильно варьироваться в зависимости от «удачности» начальных приближений.

В данной работе было реализовано две модификации метода поиска сгущений: «стандартный» и «с полным просмотром».

1. «Стандартный» вариант:
   1. Из множества рассматриваемых точек по какому-то принципу выбирается первое начальное приближение – центр кластера.
   2. Все рассматриваемые точки, лежащие «внутри» кластерного круга (т.е. расстояние до которых от центра кластера меньше радиуса), включаются в текущий кластер, все точки, лежащие вне круга, – кластеру не принадлежат.
   3. Пересчитывается центр кластера.
   4. Пункты 2 и 3 повторяются до тех пор, пока кластер не стабилизируется, т.е. нельзя ни добавить, ни исключить точки.
   5. Все точки, добавленные в кластер, исключаются из рассмотрения. Начинается новая итерация поиска кластера с пункта 1.
   6. Поиск продолжается, пока рассматриваемое множество не станет пусто.

Принципы выбора начальных приближений:

* Random – случайным образом.
* Minmedian – необходимо посчитать расстояния между всеми рассматриваемыми точками, затем для всех точки найти медианные расстояния и выбрать минимальное из них, точку с минимальным медианным расстоянием принять за начальное приближение.
* Maxmedian – то же самое, что minmedian, только выбирать максимальную медиану.

1. «С полным просмотром»:
   1. Просматриваются все точки множества: каждая точка по очереди принимается за начальное приближение и находятся N (объем выборки) кандидатов в кластеры (уникальных кластеров может оказаться меньше), как описано выше.
   2. В качестве найденного кластера выбирается содержащий наибольшее количество точек. Данные точки исключаются из дальнейшего поиска.
   3. Повторяются первый и второй пункты для оставшихся точек, пока все точки исходного множества не будут разделены на кластеры.

Для измерения расстояния используется евклидово расстояние между точками в двумерном пространстве:

За центр кластера принимается центр масс (барицентр, центроид) точек, ему принадлежащих:

**Задание.**

Дано конечное множество из объектов, представленных двумя признаками (в качестве этого множества принимаем исходную двумерную выборку, сформированную ранее в лабораторной работе №4). Выполнить разбиение исходного множества объектов на конечное число подмножеств (кластеров) с использованием метода поиска сгущений. Полученные результаты содержательно проинтерпретировать.

**Ход выполнения.**

1. Двумерная выборка объема *N = 107*.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| v | 501.00 | 369.00 | 344.00 | 473.00 | 426.00 | 528.00 | 497.00 | 467.00 | 506.00 | 431.00 | 454.00 |
| E | 130.40 | 84.30 | 86.80 | 137.90 | 121.10 | 163.40 | 147.30 | 140.50 | 158.40 | 125.00 | 131.10 |
| v | 371.00 | 482.00 | 393.00 | 441.00 | 463.00 | 440.00 | 481.00 | 340.00 | 468.00 | 397.00 | 496.00 |
| E | 89.20 | 139.90 | 103.20 | 122.80 | 129.10 | 128.50 | 135.20 | 85.10 | 142.00 | 108.60 | 143.10 |
| v | 434.00 | 541.00 | 352.00 | 438.00 | 453.00 | 423.00 | 351.00 | 525.00 | 409.00 | 469.00 | 386.00 |
| E | 122.30 | 146.80 | 87.70 | 134.90 | 119.50 | 131.10 | 89.00 | 165.90 | 121.00 | 131.50 | 95.50 |
| v | 505.00 | 436.00 | 488.00 | 449.00 | 493.00 | 512.00 | 472.00 | 423.00 | 465.00 | 351.00 | 359.00 |
| E | 137.50 | 114.30 | 134.10 | 124.50 | 129.70 | 169.90 | 134.20 | 130.80 | 140.70 | 102.90 | 71.90 |
| v | 457.00 | 467.00 | 400.00 | 418.00 | 492.00 | 434.00 | 510.00 | 392.00 | 463.00 | 459.00 | 397.00 |
| E | 126.40 | 135.10 | 114.60 | 118.60 | 137.50 | 110.50 | 140.60 | 82.70 | 125.00 | 145.40 | 106.80 |
| v | 424.00 | 436.00 | 429.00 | 398.00 | 493.00 | 522.00 | 518.00 | 463.00 | 437.00 | 386.00 | 493.00 |
| E | 119.00 | 116.70 | 112.90 | 109.00 | 154.50 | 154.50 | 144.40 | 121.20 | 121.80 | 105.80 | 151.20 |
| v | 414.00 | 480.00 | 585.00 | 562.00 | 508.00 | 421.00 | 463.00 | 422.00 | 406.00 | 544.00 | 345.00 |
| E | 113.50 | 153.90 | 177.70 | 175.90 | 159.00 | 117.80 | 136.70 | 122.90 | 110.10 | 166.70 | 95.90 |
| v | 478.00 | 393.00 | 437.00 | 448.00 | 458.00 | 422.00 | 468.00 | 430.00 | 371.00 | 543.00 | 471.00 |
| E | 126.60 | 122.80 | 115.10 | 121.90 | 121.70 | 115.70 | 144.90 | 104.30 | 91.90 | 155.40 | 143.90 |
| v | 475.00 | 521.00 | 353.00 | 437.00 | 362.00 | 490.00 | 484.00 | 459.00 | 480.00 | 482.00 | 522.00 |
| E | 132.00 | 139.60 | 98.00 | 118.40 | 111.70 | 139.90 | 140.40 | 136.70 | 153.30 | 148.20 | 143.80 |
| v | 576.00 | 390.00 | 514.00 | 442.00 | 421.00 | 443.00 | 438.00 | 429.00 |  |  |  |
| E | 166.40 | 91.40 | 153.60 | 115.40 | 107.90 | 121.90 | 126.70 | 120.90 |  |  |  |

Было произведено масштабирование значений выборки так, чтобы они все попадали в интервал [0; 1].

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| v | 0.65714286 | 0.11836735 | 0.01632653 | 0.54285714 | 0.35102041 | 0.76734694 | 0.64081633 | 0.51836735 | 0.67755102 | 0.37142857 | 0.46530612 |
| E | 0.5529301 | 0.1172023 | 0.1408318 | 0.6238185 | 0.4650284 | 0.8648393 | 0.7126654 | 0.6483932 | 0.8175803 | 0.5018904 | 0.5595463 |
| v | 0.12653061 | 0.57959184 | 0.21632653 | 0.41224490 | 0.50204082 | 0.40816327 | 0.57551020 | 0.00000000 | 0.52244898 | 0.23265306 | 0.63673469 |
| E | 0.1635161 | 0.6427221 | 0.2958412 | 0.4810964 | 0.5406427 | 0.5349716 | 0.5982987 | 0.1247637 | 0.6625709 | 0.3468809 | 0.6729679 |
| v | 0.38367347 | 0.82040816 | 0.04897959 | 0.40000000 | 0.46122449 | 0.33877551 | 0.04489796 | 0.75510204 | 0.28163265 | 0.52653061 | 0.18775510 |
| E | 0.4763705 | 0.7079395 | 0.1493384 | 0.5954631 | 0.4499055 | 0.5595463 | 0.1616257 | 0.8884688 | 0.4640832 | 0.5633270 | 0.2230624 |
| v | 0.67346939 | 0.39183673 | 0.60408163 | 0.44489796 | 0.62448980 | 0.70204082 | 0.53877551 | 0.33877551 | 0.51020408 | 0.04489796 | 0.07755102 |
| E | 0.6200378 | 0.4007561 | 0.5879017 | 0.4971645 | 0.5463138 | 0.9262760 | 0.5888469 | 0.5567108 | 0.6502836 | 0.2930057 | 0.0000000 |
| v | 0.47755102 | 0.51836735 | 0.24489796 | 0.31836735 | 0.62040816 | 0.38367347 | 0.69387755 | 0.21224490 | 0.50204082 | 0.48571429 | 0.23265306 |
| E | 0.5151229 | 0.5973535 | 0.4035917 | 0.4413989 | 0.6200378 | 0.3648393 | 0.6493384 | 0.1020794 | 0.5018904 | 0.6947070 | 0.3298677 |
| v | 0.34285714 | 0.39183673 | 0.36326531 | 0.23673469 | 0.62448980 | 0.74285714 | 0.72653061 | 0.50204082 | 0.39591837 | 0.18775510 | 0.62448980 |
| E | 0.4451796 | 0.4234405 | 0.3875236 | 0.3506616 | 0.7807183 | 0.7807183 | 0.6852552 | 0.4659735 | 0.4716446 | 0.3204159 | 0.7495274 |
| v | 0.30204082 | 0.57142857 | 1.00000000 | 0.90612245 | 0.68571429 | 0.33061224 | 0.50204082 | 0.33469388 | 0.26938776 | 0.83265306 | 0.02040816 |
| E | 0.3931947 | 0.7750473 | 1.0000000 | 0.9829868 | 0.8232514 | 0.4338374 | 0.6124764 | 0.4820416 | 0.3610586 | 0.8960302 | 0.2268431 |
| v | 0.56326531 | 0.21632653 | 0.39591837 | 0.44081633 | 0.48163265 | 0.33469388 | 0.52244898 | 0.36734694 | 0.12653061 | 0.82857143 | 0.53469388 |
| E | 0.5170132 | 0.4810964 | 0.4083176 | 0.4725898 | 0.4706994 | 0.4139887 | 0.6899811 | 0.3062382 | 0.1890359 | 0.7892250 | 0.6805293 |
| v | 0.55102041 | 0.73877551 | 0.05306122 | 0.39591837 | 0.08979592 | 0.61224490 | 0.58775510 | 0.48571429 | 0.57142857 | 0.57959184 | 0.74285714 |
| E | 0.5680529 | 0.6398866 | 0.2466919 | 0.4395085 | 0.3761815 | 0.6427221 | 0.6474480 | 0.6124764 | 0.7693762 | 0.7211720 | 0.6795841 |
| v | 0.96326531 | 0.20408163 | 0.71020408 | 0.41632653 | 0.33061224 | 0.42040816 | 0.40000000 | 0.36326531 |  |  |  |
| E | 0.8931947 | 0.1843100 | 0.7722117 | 0.4111531 | 0.3402647 | 0.4725898 | 0.5179584 | 0.4631380 |  |  |  |

Теперь обе величины, составляющие двумерную выборку, вносят одинаковый вклад при расчете расстояния.

1. В работе было проведено два эксперимента:
   * + 1. в первом при построении кластеров варьировался радиус и использовались различные модификации алгоритма (выбор начального приближения),
       2. во втором эксперименте фиксировался радиус, а начальные приближения выбирались случайно.

**Результаты первого эксперимента.**

Радиус варьировался в промежутке [0.1; 0.325] c шагом в δ = 0.025.

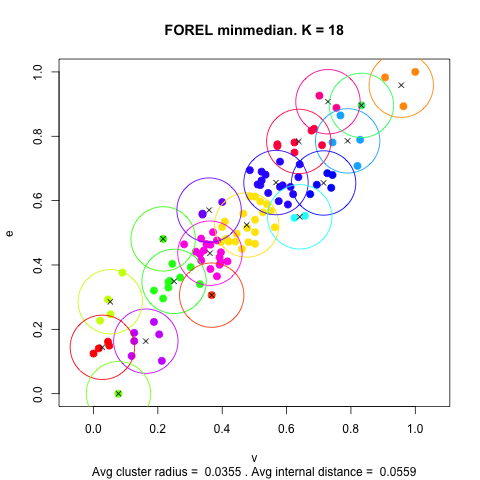
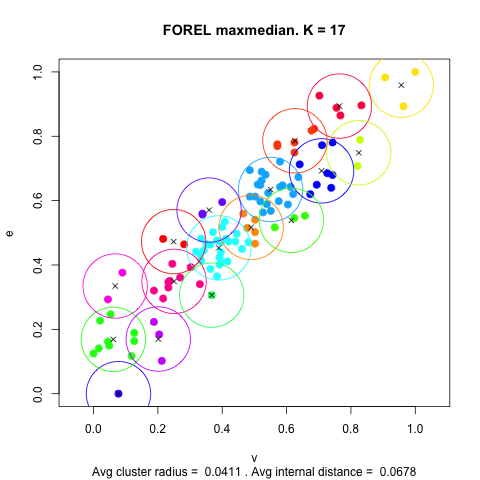
K – количество найденных кластеров.

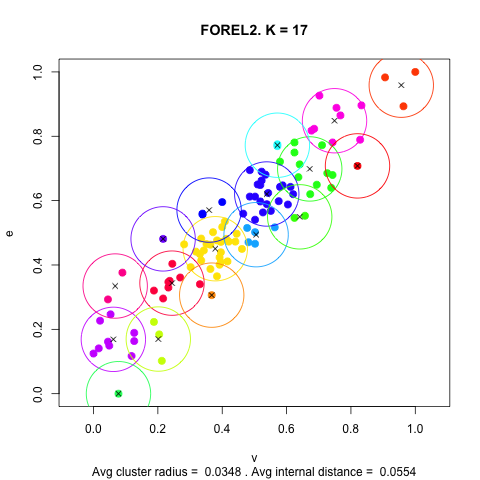
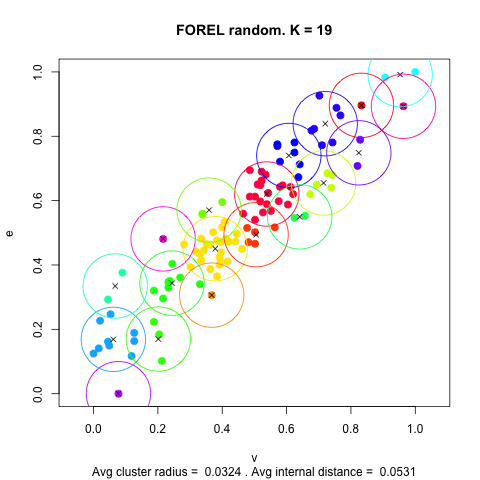
FOREL maxmedian - вариант алгоритма, при котором в качестве начального приближения выбирается точка, имеющую максимальное значение медианы (среднее расстояние до остальных точек),

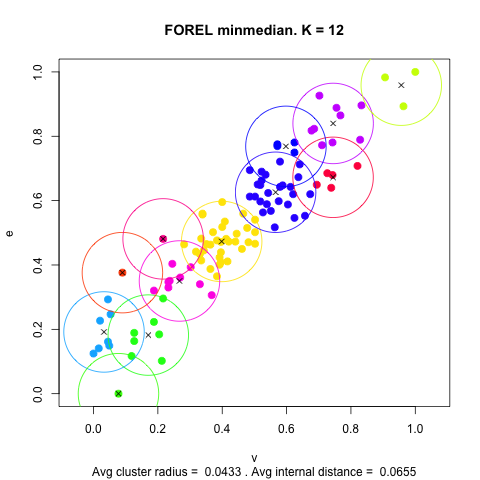
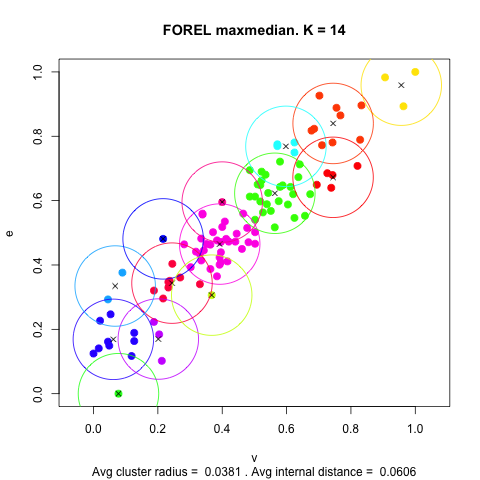
FOREL minmedian - вариант алгоритма, при котором в качестве начального приближения выбирается точка, имеющую минимальное значение медианы (среднее расстояние до остальных точек),

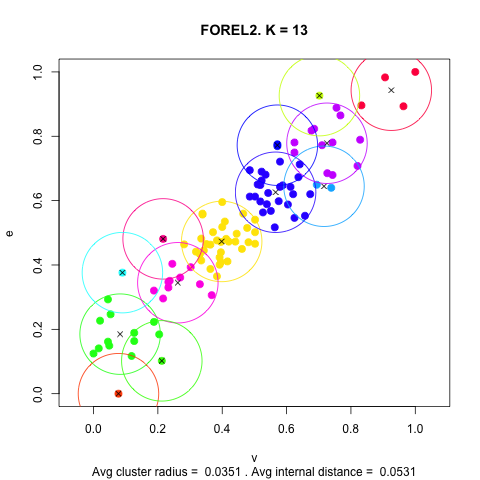
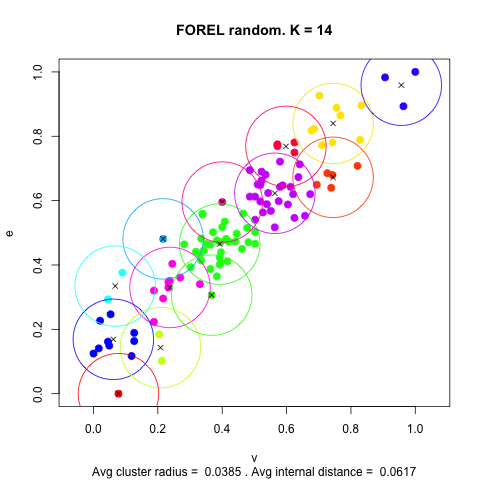
FOREL random – вариант алгоритма с соответствующим способом выбора начальных приближений.

FOREL2 – модификация алгоритма «с полным просмотром».

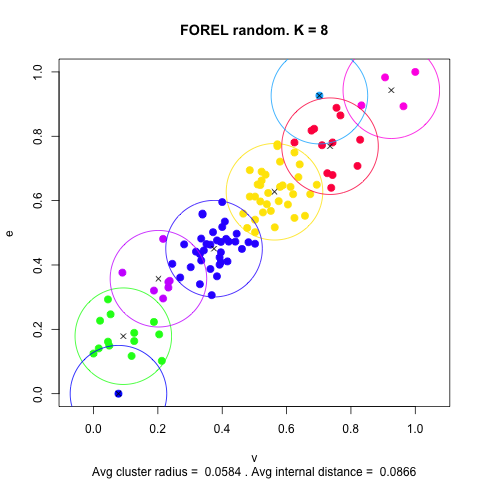
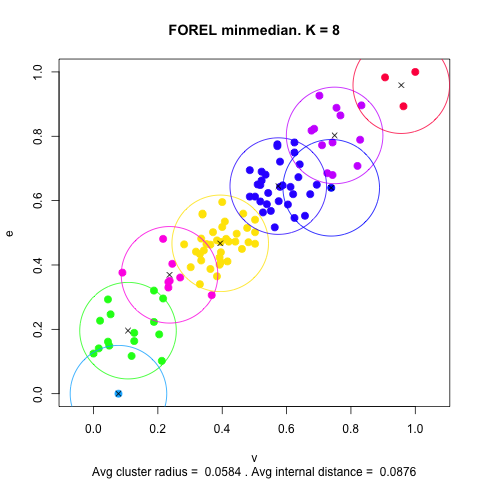
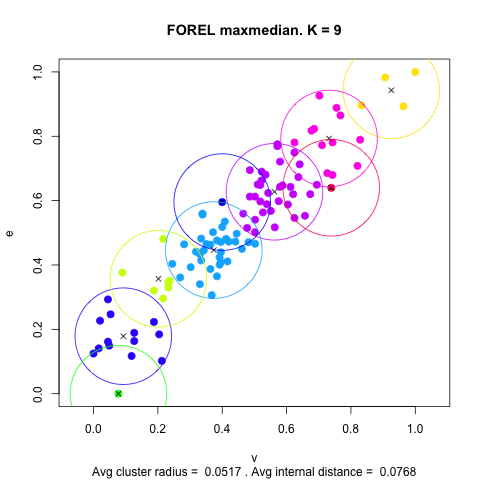
**R = 0.1**

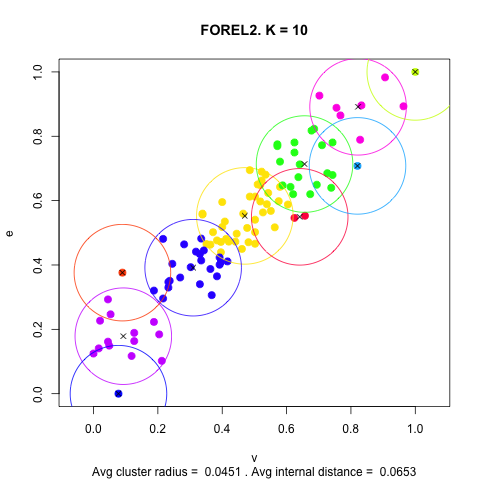
****

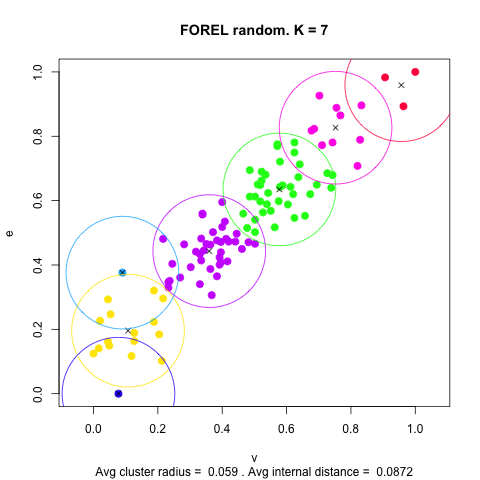
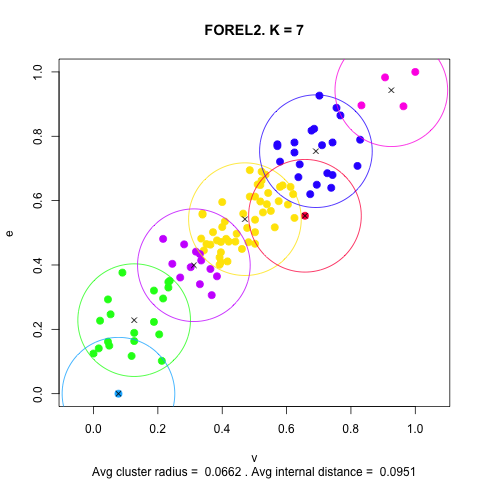
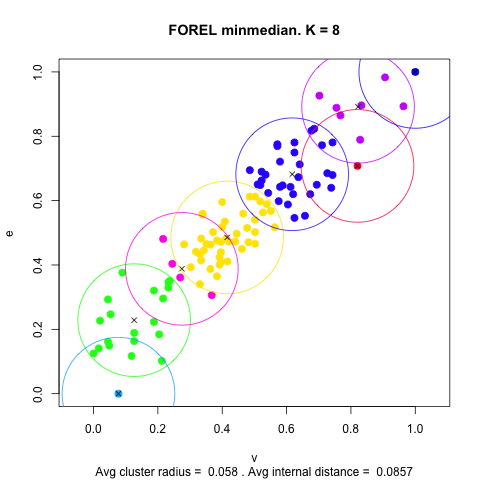
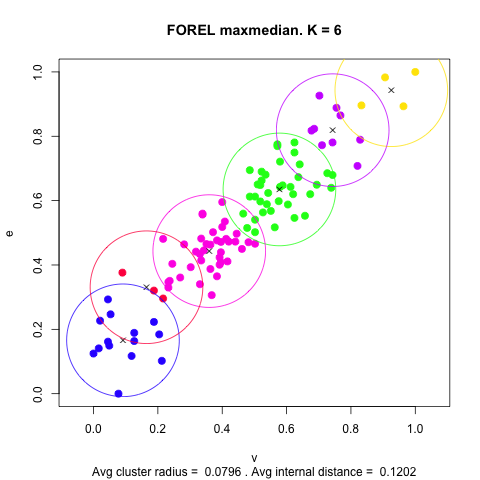
**R = 0.125**

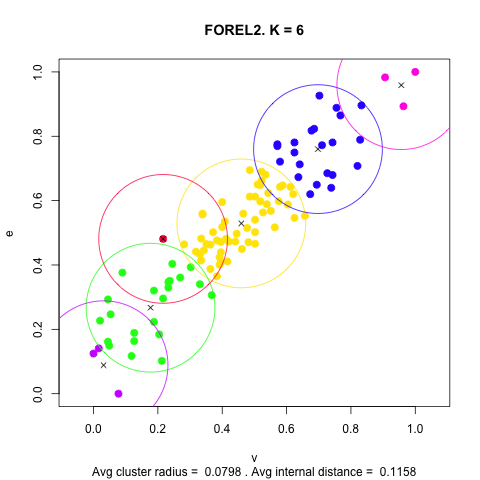
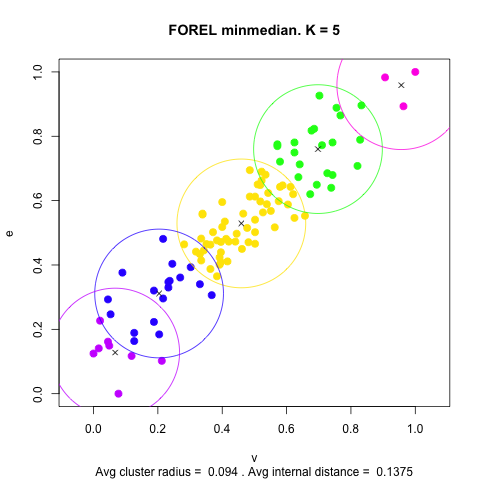
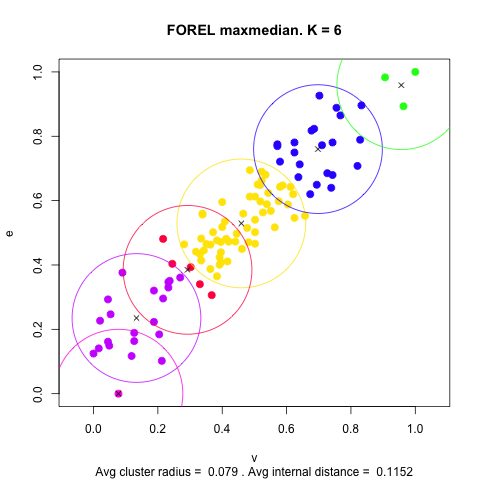
****

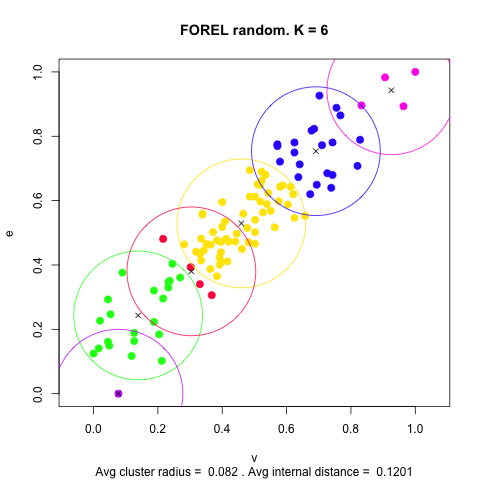
**R = 0.15**

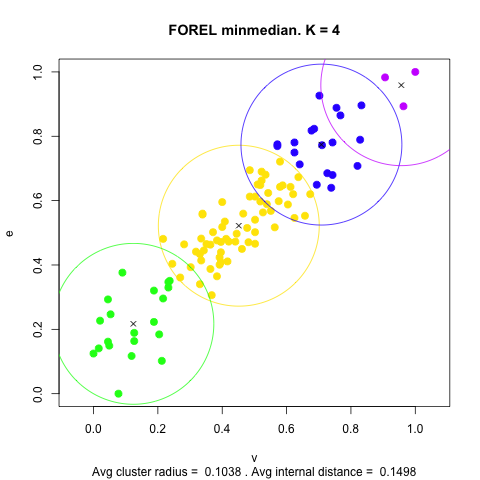
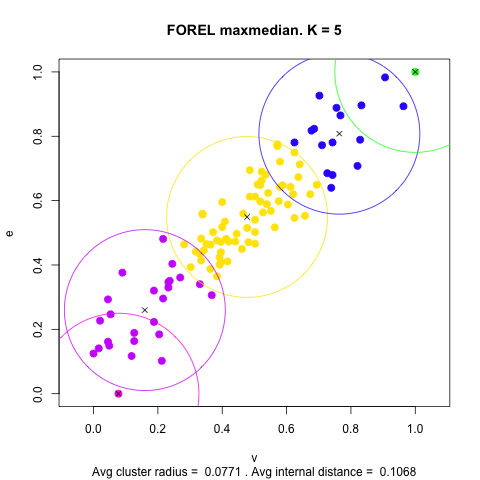
****

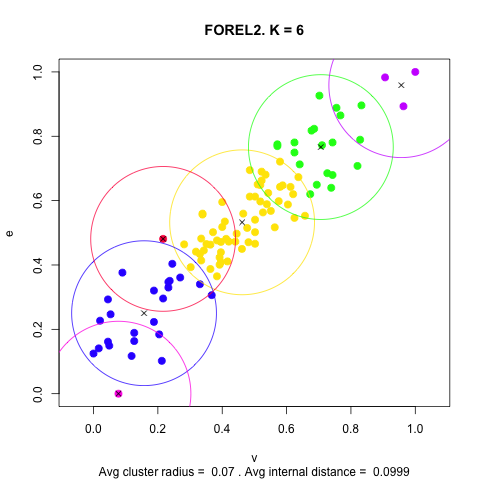
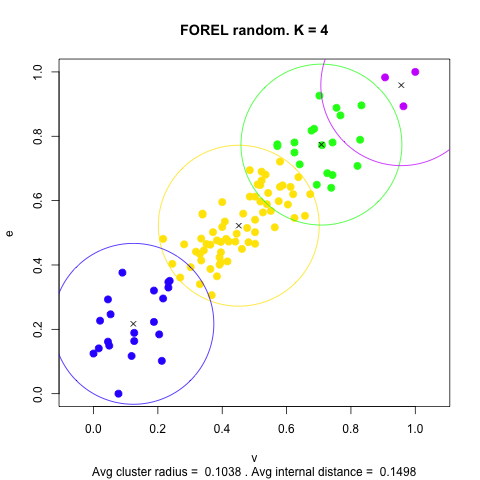


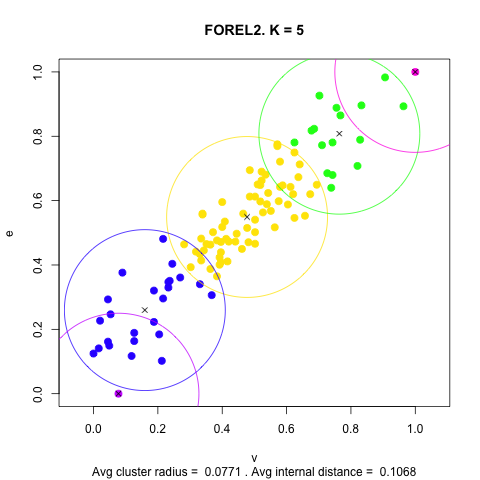
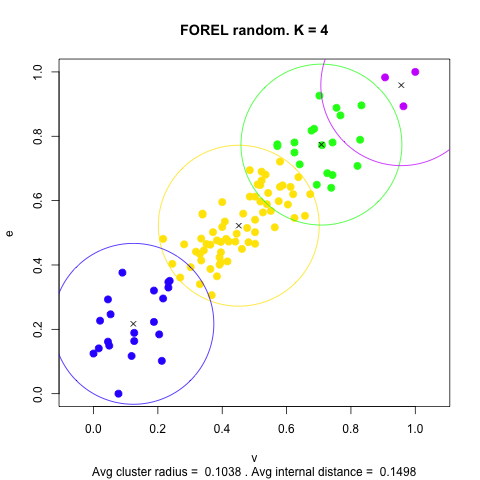
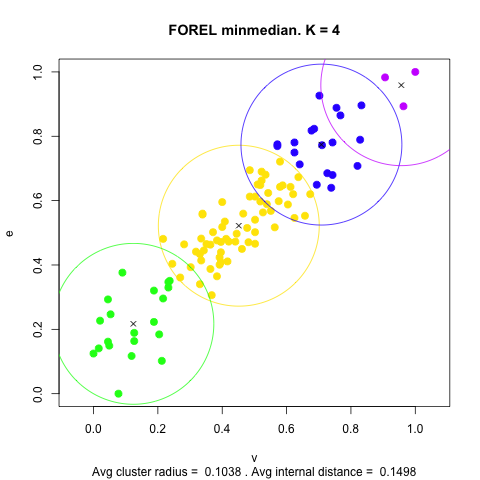
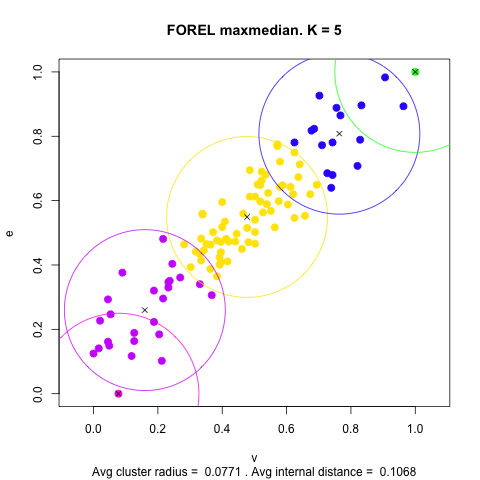
**R = 0.175**

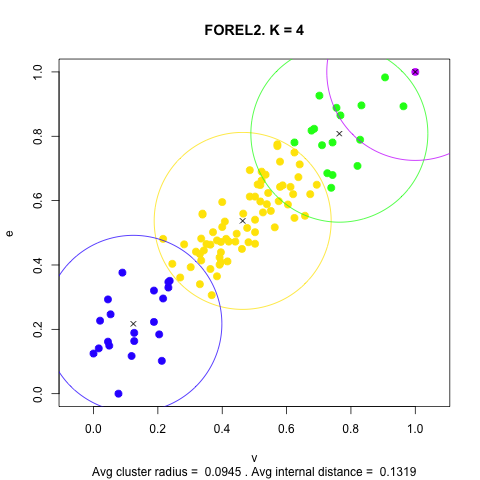
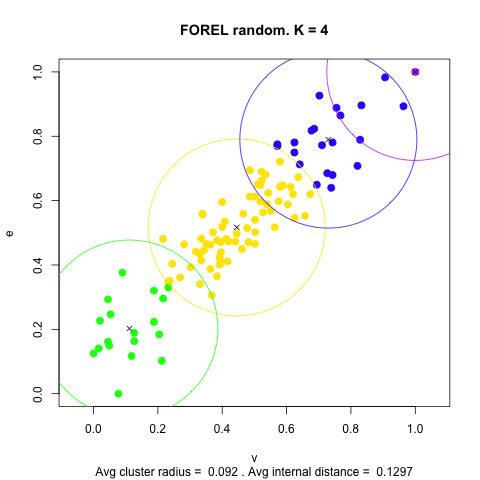
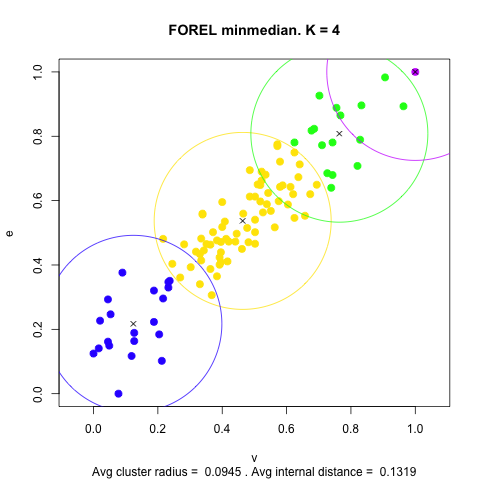
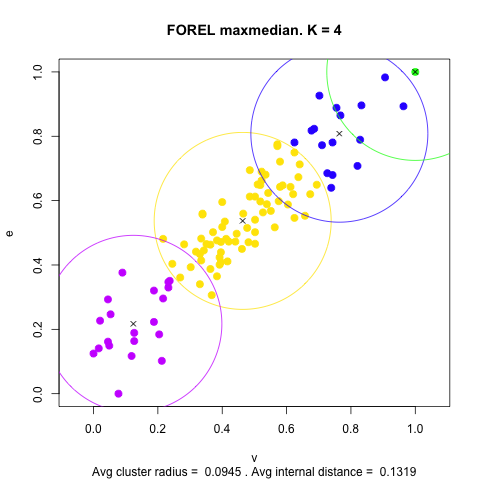
**R = 0.2**

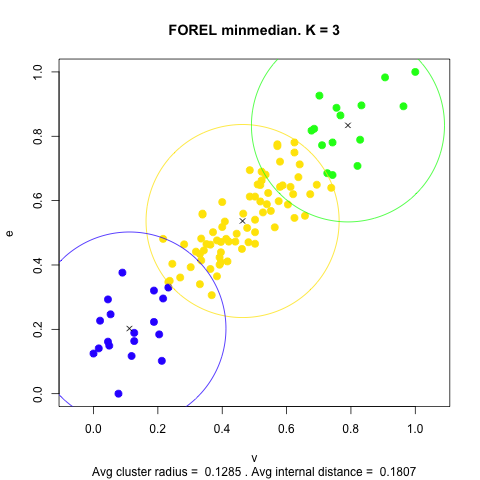
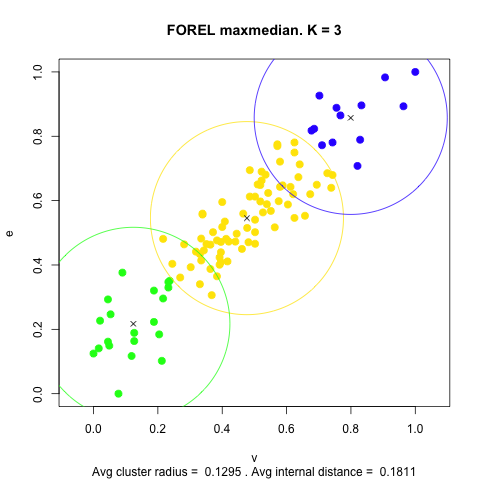
****

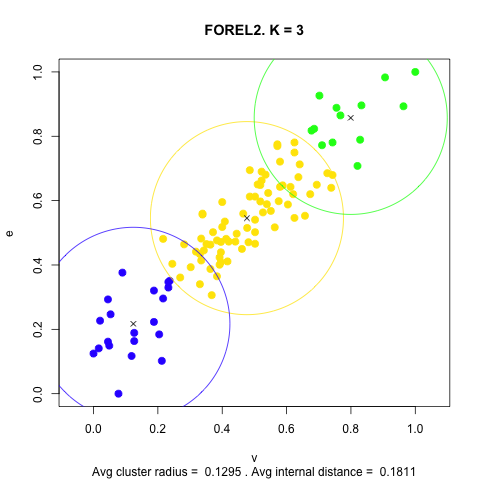
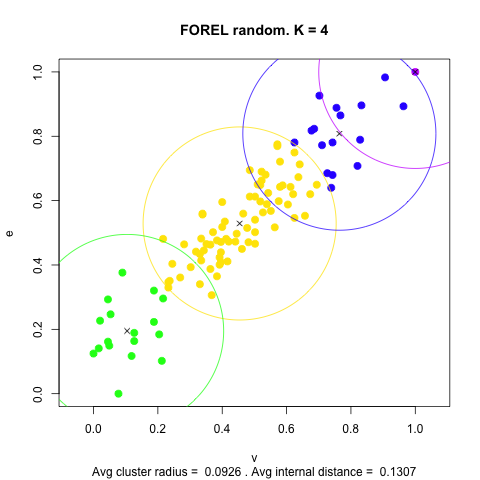
**R = 0.225**

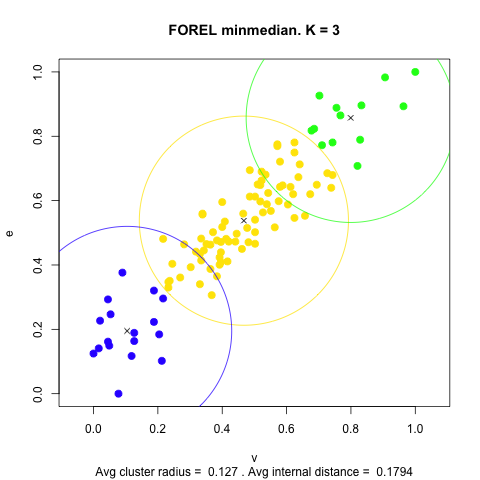
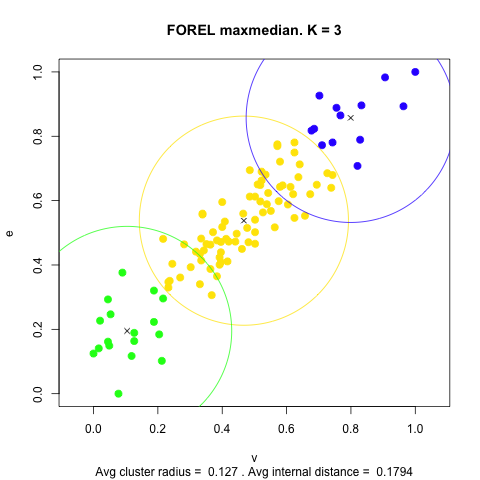
****

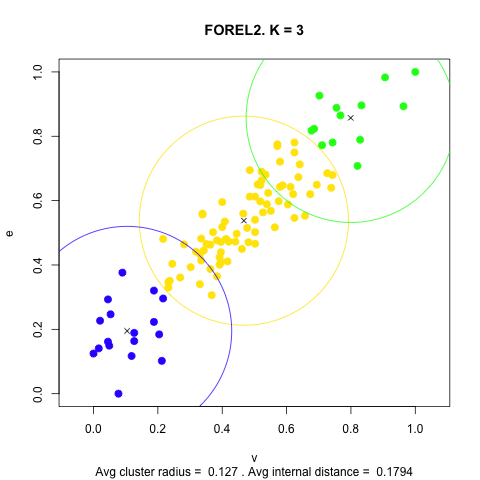
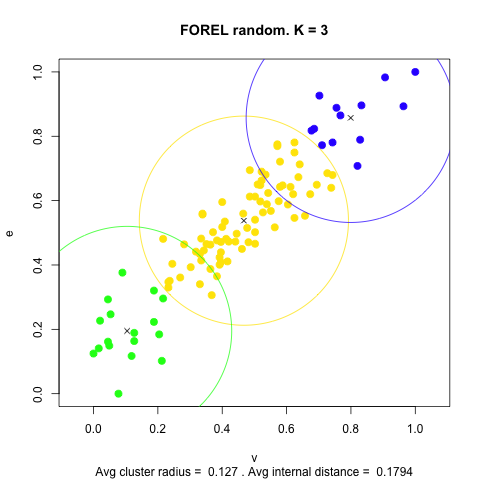
**R = 0.25**

**R = 0.275**

**R = 0.3**



**R = 0.325**

****

Сравнительная таблица

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| R | Средний кластерный радиус | | | | Среднее внутрикластерное расстояние | | | |
| Стандартный | | | С полным просмотром | Стандартный | | | С полным просмотром |
| Maxmedian | Minmedian | Random | Maxmedian | Minmedian | Random |
| 0.1 | 0.04109562 | 0.03553614 | 0.03236344 | 0.03479994 | 0.06781942 | 0.05594195 | 0.05310429 | 0.05536840 |
| 0.125 | 0.03812300 | 0.04327640 | 0.03848928 | 0.03512750 | 0.06058621 | 0.06551348 | 0.06169720 | 0.05311393 |
| 0.15 | 0.05170254 | 0.05835503 | 0.05841075 | 0.04506971 | 0.07675329 | 0.08758054 | 0.08659528 | 0.06531378 |
| 0.175 | 0.07960202 | 0.05803598 | 0.05900465 | 0.06623173 | 0.12023622 | 0.08565217 | 0.08718532 | 0.09511398 |
| 0.2 | 0.07902859 | 0.09402532 | 0.08197344 | 0.07981205 | 0.11520496 | 0.13749269 | 0.12009480 | 0.11583243 |
| 0.225 | 0.06997255 | 0.06997255 | 0.05938874 | 0.06997255 | 0.09987377 | 0.09987377 | 0.08471683 | 0.09987377 |
| 0.25 | 0.07714720 | 0.10379381 | 0.10379381 | 0.07714720 | 0.10676493 | 0.14981927 | 0.14981927 | 0.10676493 |
| 0.275 | 0.09445128 | 0.09445128 | 0.09201983 | 0.09445128 | 0.13185564 | 0.13185564 | 0.12969719 | 0.13185564 |
| 0.3 | 0.12945532 | 0.12850745 | 0.09258588 | 0.12945532 | 0.18105053 | 0.18074903 | 0.13066532 | 0.18105053 |
| 0.325 | 0.12699336 | 0.12699336 | 0.12699336 | 0.12699336 | 0.17940961 | 0.17940961 | 0.17940961 | 0.17940961 |
| R | Количество кластеров | | | |  | | | |
| Стандартный | | | С полным просмотром |  | | | |
| Maxmedian | Minmedian | Random |  | | | |
| 0.1 | 17 | 18 | 19 | 17 |  | | | |
| 0.125 | 14 | 12 | 14 | 13 |  | | | |
| 0.15 | 9 | 8 | 8 | 10 |  | | | |
| 0.175 | 6 | 8 | 7 | 7 |  | | | |
| 0.2 | 6 | 5 | 6 | 6 |  | | | |
| 0.225 | 6 | 6 | 7 | 6 |  | | | |
| 0.25 | 5 | 4 | 4 | 5 |  | | | |
| 0.275 | 4 | 4 | 4 | 4 |  | | | |
| 0.3 | 3 | 3 | 4 | 3 |  | | | |
| 0.325 | 3 | 3 | 3 | 3 |  | | | |

Для заданного радиуса в зависимости от способа выбора начальных приближений получается различное количество кластеров, в силу этого сложно оценивать качество полученных кластеров с помощью рассматриваемых характеристик.

Например, если рассмотреть три нижние строки в таблице, представленной выше, можно отметить, что, во-первых, количество полученных кластеров для всех четырех способов одинаково и равно трем, во-вторых, наилучший результат показала модификация «с полным просмотром» при заданном радиусе R = 0.325.



* 1. Результаты второго эксперимента.

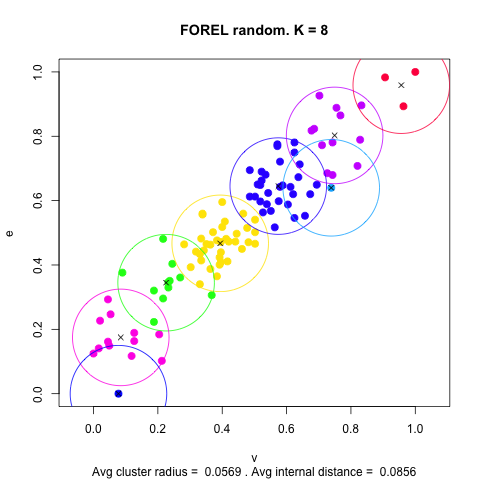
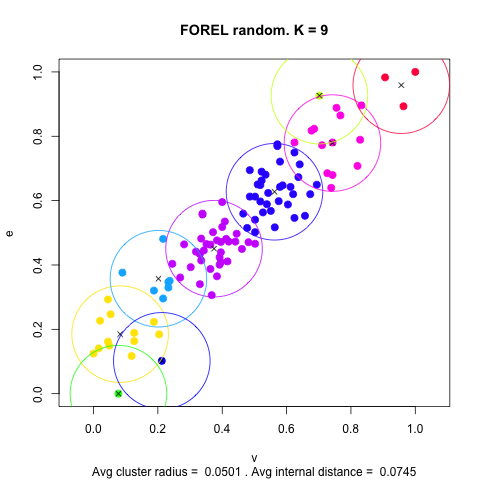
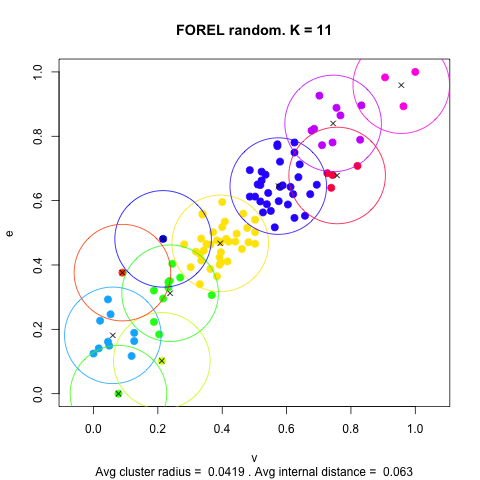
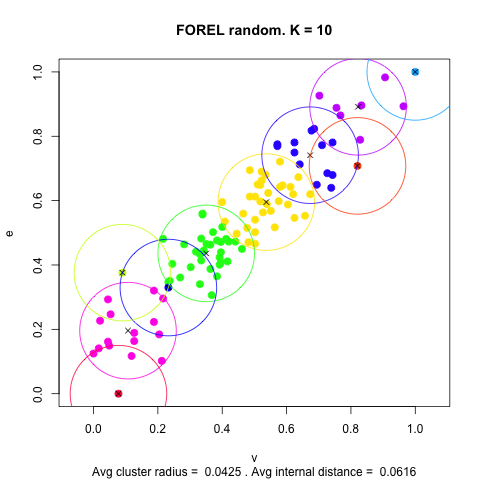
Значение радиуса было зафиксировано на уровне *R = 0.15*.

В этом эксперименте стандартный вариант алгоритма был запущен 10 раз. В результате получилось 10 различных разбиений на кластеры. Начальные приближения выбирались случайно. Далее представлена таблица с характеристиками разбиения.

Получившиеся разбиения можно поделить на группы по количеству выделенных кластеров (*K*). На графиках представлены разбиения с лучшими характеристиками в своей группе.

Сравнительная таблица.

|  |  |  |  |
| --- | --- | --- | --- |
| № | Средний кластерный радиус | Среднее внутрикластерное расстояние | Количество кластеров |
| 1 | 0.05252209 | 0.07770190 | 9 |
| 2 | 0.05382640 | 0.07967489 | 9 |
| **3** | **0.04245416** | **0.06156258** | **10** |
| 4 | 0.05841075 | 0.08659528 | 8 |
| 5 | 0.05054584 | 0.07553307 | 9 |
| 6 | 0.05072532 | 0.07607701 | 9 |
| 7 | 0.05087707 | 0.07609360 | 9 |
| **8** | **0.04185156** | **0.06298835** | **11** |
| **9** | **0.05013994** | **0.07454487** | **9** |
| **10** | **0.05688853** | **0.08563213** | **8** |



**Выводы.**

В ходе лабораторной работы был реализован алгоритм поиска сгущений (ForEl) для кластеризации точек на двумерной плоскости в двух модификациях: «стандартной» и «с полным просмотром».

Для «стандартного» варианта было реализовано три способа выбора начальных приближений.

Было проведено два набора экспериментов: в первом варьировался радиус и способы поиска кластеров сравнивались между собой, во втором для фиксированного радиуса и случайных начальных приближений несколько раз запускался «стандартный» алгоритм для определения надежности характеристик полученного разбиения.

По результатам проведенного исследования можно сделать вывод, что при использовании метода поиска сгущений, значения среднего кластерного расстояния и среднего кластерного радиуса достигаются в случае успешного нахождения плотных зон (сгущений). Однако, при таком разбиении остаются точки, образующие кластеры нулевого радиуса и с нулевым кластерным расстоянием.

Также полученные значения среднего внутрикластерного расстояния и среднего кластерного радиуса довольно устойчивы, при фиксации радиуса кластера.

Основной задачей работы является визуализация результатов работы алгоритма. Результаты представлены в тексте работы в виде графиков и таблиц.