

5640

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
РЯЗАНСКИЙ ГОСУДАРСТВЕННЫЙ РАДИОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. В.Ф. УТКИНА

КЛАСТЕРИЗАЦИЯ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМА k-СРЕДНИХ

Методические указания к лабораторной работе

Рязань 2021

УДК 004.825

Кластеризация данных с использованием алгоритма k-средних: методические указания к лабораторной работе / Рязан. гос. радиотехн. ун-т; сост.: В.И. Орешков, И.А. Цепулин. Рязань, 2021. 16 с.

Рассмотрены теоретические и практические основы алгоритма кластеризации k-средних. Изучен процесс реализации алгоритма кластеризации k-средних в аналитической платформе Deductor Academic.

Предназначены для магистрантов направления 09.04.01 «Информатика и вычислительная техника».

Табл. 2. Ил. 11. Библиогр.: 4 назв.

Кластеризация, кластер, машинное обучение, центроид

Печатается по решению редакционно-издательского совета Рязанского государственного радиотехнического университета.

Рецензент: кафедра САПР вычислительных средств Рязанского государственного радиотехнического университета (зав. кафедрой д-р техн. наук, проф. В.П. Корячко)

Кластеризация данных с использованием алгоритма k-средних

Составители: О р е ш к о в Вячеслав Игоревич
Ц е п у л и н Иван Андреевич

Редактор М.Е. Цветкова

Корректор С.В. Макушина

Подписано в печать 12.01.21. Формат бумаги 60×84 1/16.

Бумага писчая. Печать трафаретная. Усл. печ. л. 1,0.

Тираж 25 экз. Заказ

Рязанский государственный радиотехнический университет.

390005, Рязань, ул. Гагарина, 59/1.

Редакционно-издательский центр РГРТУ.

Цель работы: изучить теоретические основы и практическую реализацию алгоритма кластеризации k -средних средствами аналитической платформы Deductor.

Теоретическая часть

Общая постановка задачи кластеризации. Кластеризация является одной из наиболее важных задач анализа данных.

Кластеризация - группировка объектов (точнее, их векторов в пространстве признаков) в обособленные группы, называемые *кластерами*. При этом в каждый кластер группируются объекты с близкими значениями признаков. На первый взгляд, задачи классификации и кластеризации похожи: и в той, и в другой имеет место группировка «похожих» объектов. Но на самом деле между этими задачами есть принципиальное отличие: при классификации классы должны быть предварительно определены и описаны, а для кластеризации алгоритм должен самостоятельно объединить все представленные объекты в кластеры исключительно на основе каких-либо критериев их близости.

Кластеризация оказывается особенно полезной, когда отсутствуют априорные сведения о структурных характеристиках набора данных. Ещё одним важным моментом является отсутствие необходимости использовать в кластерных моделях целевую переменную, поэтому для построения кластерной модели может использоваться обучение без учителя.

С точки зрения извлечения новых знаний кластеризация представляет больший интерес, чем классификация. Действительно, при классификации любой объект, предъявляемый модели, обязательно будет отнесен к одному из заранее определённых классов, даже если в действительности он ни к одному из классов не относится. В случае кластеризации «нетипичные» объекты образуют новый кластер, что позволяет обнаружить их новые свойства. Иными словами, кластеризация более полезна с точки зрения «обнаружения новизны», чем классификация, поскольку позволяет обнаруживать и описывать объекты с ранее неизвестными, новыми свойствами.

Решение задачи кластеризации обычно сложнее, чем классификации, поскольку число кластеров, как правило, априори неизвестно, а выбор меры «близости» объектов в пространстве признаков и критерия качества кластеризации часто носит субъективный характер. В методах машинного обучения объекты и наблюдения представляются в виде векторов в многомерном пространстве, расстояния между которыми определяются в соответствии с некоторой метрикой (Евклида, городских кварталов и т.д.). Поэтому в контексте интеллектуального

анализа данных можно определить кластер как группу векторов в пространстве признаков такую, что расстояние между двумя любыми векторами внутри кластера всегда меньше, чем расстояние до любого вектора из другого кластера (рис. 1).

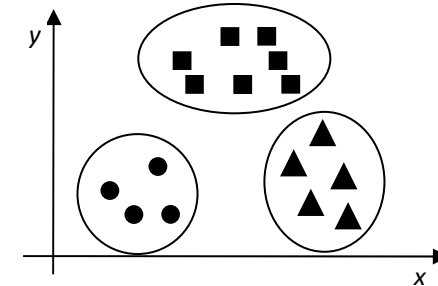


Рис. 1. Представление кластеров в 2-мерном случае

Также кластер можно определить, как область векторного пространства, содержащую группу векторов, удовлетворяющих приведенному выше условию.

Этапы кластерного анализа. Кластерный анализ данных состоит из 4-х шагов.

1. Построение кластерной модели на обучающем наборе.
2. Содержательная интерпретация кластеров – изучение общих свойств объектов, попавших в каждый кластер.
3. Оценка состоятельности модели.
4. Практическое использование модели: модели предъявляются новые наблюдения, которые она относит к одному из кластеров. Поскольку свойства кластеров предварительно изучены, то они могут быть обобщены для любого нового объекта, попавшего в кластер.

В настоящее время разработано большое количество алгоритмов кластеризации. Их можно разделить на иерархические и неиерархические. Иерархические алгоритмы предполагают возможность построения дерева кластерной структуры, где кластеры представляют собой «листья». Неиерархические алгоритмы не предполагают разделения кластеров на иерархические уровни.

Кроме этого, алгоритмы кластеризации разделяют на агломеративные и конгломеративные. Агломеративные алгоритмы работают путём объединения меньших кластеров в большие. Конгломеративные - посредством разделения исходного набора данных на всё меньшие группы. Процесс формирования кластеров идёт до тех пор, пока алгоритмом не будет достигнуто оптимальное значение некоторой целевой функции.

Теоретическая основа алгоритма кластеризации k-средних

Идея алгоритма k-means (k-средних) была одновременно сформулирована Г. Штейнгаузом и С. Ллойдом в 1957 г. Сам термин «k-средних» был впервые введен Дж. Маккуинном в 1967 г. Алгоритм относится к методам машинного обучения и используется для решения задач кластеризации. В основе алгоритма лежит итеративное разбиение векторного пространства на заранее заданное число кластеров k. Алгоритм получил широкое распространение благодаря относительной простоте, наглядности реализации и достаточно высокой точности.

Пусть набор данных X содержит n наблюдений x_i . Тогда можно записать:

$$X = (x_1, x_2, \dots, x_n), i = 1, n.$$

Число кластеров k, на которое требуется разбить X: $k \in \mathbb{N}, k < n$. Тогда требуется разбить X на k кластеров C_1, C_2, \dots, C_k таким образом, что:

1) $C_i \cap C_j = \emptyset, i \neq j$, то есть кластеры не пересекаются (каждое наблюдение может принадлежать только одному кластеру);

2) $\bigcup_{i=1}^k C_i = X$ то есть в кластеры должны быть распределены все наблюдения из X.

Алгоритм разбивает исходное множество наблюдений на k кластеров C_1, C_2, \dots, C_k , при этом целью итеративного процесса является минимизация суммы квадратов расстояний от вектора каждого наблюдения до центра (центроида, центра масс) кластера:

$$\arg \min \sum_{i=1}^k \sum_{x \in C_i} d(x, \mu_i)^2,$$

где x - вектор наблюдения; μ_i - центроид i-го кластера; $d(x, \mu_i)$ - функция расстояния. При этом

$$\mu_i = \frac{1}{N_i} \sum_{x \in C_i} x,$$

$$d(v_1, v_2) = \sqrt{\sum_{i=1}^N (v_{1,i} - v_{2,i})^2}.$$

Алгоритм представляет собой итерационную процедуру, в которой выполняются следующие шаги.

1. Выбирается число кластеров k, которое в общем случае заранее неизвестно. Оно может быть выбрано случайным образом, а в дальнейшем корректироваться экспериментально, либо на основе некоторых априорных сведений о структуре исходного набора данных, либо исходя из логики предметной области.

2. Инициализация кластеров. Из исходного множества данных случайным образом выбираются k наблюдений, которые станут начальными центрами кластеров. Такие начальные центры кластеров иногда называют «семенами» или «зародышами».

3. Для каждого наблюдения исходного набора определяется ближайший к нему центр кластера (расстояния измеряются в метрике Евклида). При этом записи, «притянутые» определенным центром, образуют начальные кластеры.

4. Вычисляются центроиды — центры тяжести кластеров. Каждый центроид — это вектор, элементы которого представляют собой средние значения соответствующих признаков, вычисленные по всем записям кластера.

5. Центр кластера смещается в его центроид, после чего центроид становится центром нового кластера.

6. Шаги 3 - 5 итеративно повторяются. Очевидно, что на каждой итерации происходит изменение границ кластеров и, следовательно, смещение их центров, которые вычисляются заново. В результате минимизируется расстояние между элементами внутри кластеров и увеличиваются междукластерные расстояния.

Остановка алгоритма производится тогда, когда границы кластеров и расположения центроидов не перестанут изменяться от итерации к итерации, то есть на каждой итерации в каждом кластере будет оставаться один и тот же набор наблюдений. На практике алгоритм обычно находит набор стабильных кластеров за несколько десятков итераций.

К недостаткам алгоритма можно отнести неопределенность выбора начальных центров и числа кластеров.

Существуют методы кластеризации, которые можно рассматривать как происходящие от k-средних. Например, в методе k-медиан (k-medoids) для вычисления центроидов используется не среднее, а медиана, что делает алгоритм более устойчивым к аномальным значениям в данных.

Алгоритм g-средних (от gaussian) строит кластеры, распределение данных в которых стремится к нормальному и снимает неопределенность выбора начальных кластеров.

Алгоритм c-средних (англ. fuzzy clustering, soft k-means, c-means) использует элементы нечеткой логики, учитывая при вычислении центроидов не только расстояния, но и степень принадлежности наблюдения к

множеству объектов в кластере. Метод нечеткой кластеризации *c*-средних имеет ограниченное применение из-за существенного недостатка — невозможности корректного разбиения в случае, когда кластеры имеют различную дисперсию по различным размерностям (осям) элементов (например, кластер имеет форму эллипса).

Также известен алгоритм Ллойда, который в качестве начального разбиения использует не множества векторов, а области векторного пространства.

1. Кластеризация данных с помощью алгоритма *k-means* в аналитической платформе *Deductor*

1. Загрузить в аналитическое приложение исходный набор данных по указанию преподавателя.

2. Открыть *Мастер обработки* и в секции *Data Mining* выбрать пункт *Кластеризация - кластеризация алгоритмом k-means* (рис. 2) и нажать *Далее*.

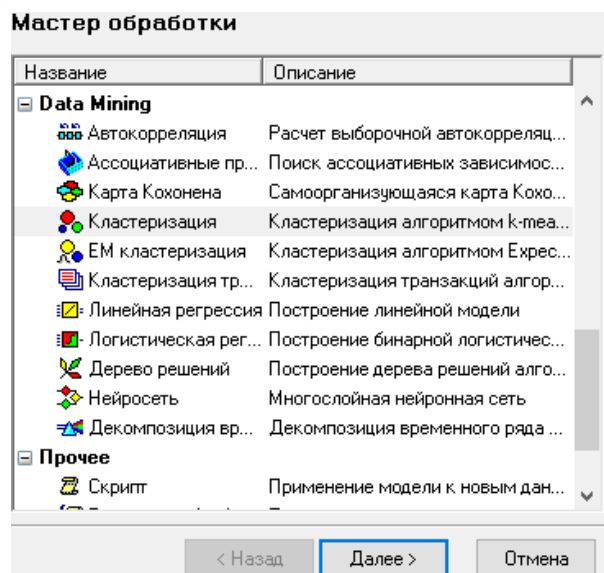


Рис. 2. Выбор алгоритма кластеризации *k-means*

2. Настройка назначения полей исходного набора данных.

Назначение «Неиспользуемое» позволяет исключить поле из построения кластеризационной модели.

Информационное поле не используется в процессе построения модели, но будет отображаться в результирующем наборе данных, полученном в результате кластеризации. Оно позволит помочь выполнить содер-

жательную интерпретацию кластеров. Входные поля будут содержать признаки, на основе которых будет производиться кластеризация.

В качестве входных полей имеет смысл выбирать только те, на основе которых можно различать наблюдения и которые отражают логику предметной области. Например, не имеет смысла использовать поля, которые для всех наблюдений содержат одинаковые значения. Не следует также использовать поля, содержащие порядковые номера или коды наблюдений, поскольку они не отражают зависимости предметной области (рис. 3).

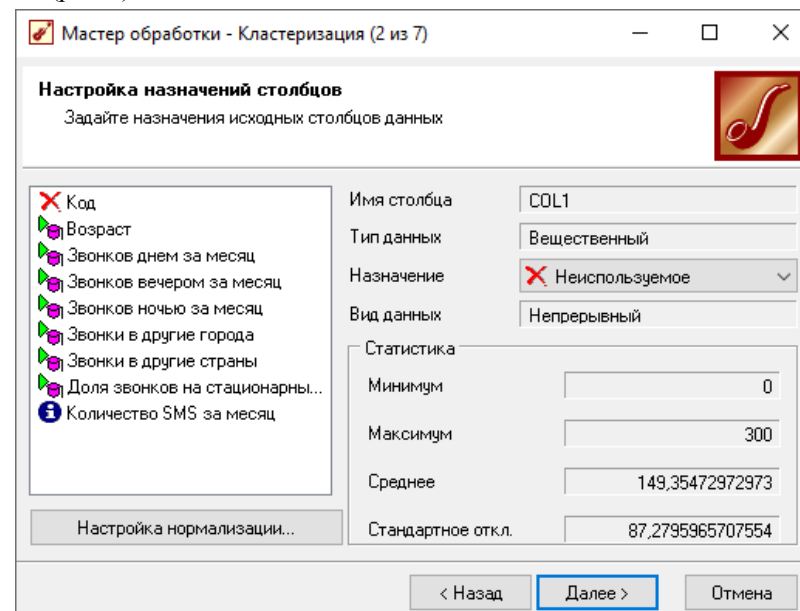


Рис. 3. Выбор назначения полей исходного набора данных

Кроме этого, следует учитывать, что хотя большее число признаков, используемых при кластеризации, как правило, повышает её точность, число признаков, большее 5 – 7, делают кластерную структуру трудно понимаемой человеком. После выбора назначений полей нажать *Далее*.

3. *Разделение исходного набора наблюдений на обучающее и тестовое множества* (рис. 4). На примерах обучающего множества будет производиться кластеризация, а на примерах тестового множества - оцениваться её точность. Если модель покажет высокую точность как на обучающем, так и на тестовом множестве, то модель обучилась хорошо. Если низкая ошибка получена только на обучающем множестве, а на тестовом -

высокая, то модель, скорее всего, переобучена и будет плохо работать на новых (практических) данных.

Разбиение исходного набора данных на подмножества
Настройте разбиение исходного множества данных на обучающее и тестовое множества

Способ разделения исходного множества данных: Случайно

Столбец для разделения исходного множества:

Множество	Размер		Порядок сортировки
	В процентах	В строках	
<input checked="" type="checkbox"/> Обучающее	95,00	281	По возрастанию
<input checked="" type="checkbox"/> Тестовое	5,00	15	По возрастанию
ИТОГО:	100,00	296	

Количество строк (всего): 296

< Назад **Далее >** Отмена

Рис. 4. Разделение исходного набора данных на обучающее и тестовое множества

Размеры обучающего и тестового множеств могут быть заданы прямым указанием числа примеров или их процентной доли от общего количества. Разделение на множества может производиться в случайном порядке (рекомендуется) и последовательно, в порядке убывания или возрастания значений некоторого признака.

Обучающее множество должно быть всегда. В то же время тестовое множество может не использоваться вообще, и, чтобы отказаться от его использования, достаточно сбросить соответствующий флажок. Отказ от использования тестового множества может быть произведён, если, например, наблюдений обучающего множества едва достаточно для обучения и отбор части примеров в тестовое множество ухудшит его результаты. На практике доля тестовых примеров в исходном наборе данных выбирается около 10 %.

После того, как разбиение на обучающее и тестовое множества настроено, следует нажать *Далее*.

4. Настройка параметров кластеризации. На данном шаге требуется выбрать способ определения числа кластеров (рис. 5).

Настройка параметров кластеризации
Укажите параметры кластеризации

☒ **Автоматически определить количество кластеров (g-means)**
Уровень значимости, %: 0,1
Чем больше уровень значимости, тем больше кластеров будет сгенерировано при кластеризации.

☐ **Фиксированное количество кластеров (k-means)**
Количество кластеров: 7
Каждая строка исходного набора данных будет отнесена к одному из кластеров, количество которых определяется выше.

< Назад **Далее >** Отмена

Рис. 5. Настройка параметров кластеризации

Автоматически выбрать число кластеров - число кластеров, на которое будет разбит исходный набор наблюдений, будет определяться автоматически, на основе заданного уровня значимости. Чем выше уровень значимости, тем больше кластеров сформирует алгоритм. Параметр определяет, насколько распределение значений в кластере согласуется с нормальным (Гауссовым) распределением. Уровень значимости задаёт порог проверочной статистики согласования распределений, при котором формирование кластера прекращается.

Такая модификация алгоритма *k-means* известна как *g-means* (*g* - от слова «гаусс») и позволяет автоматизировать процесс выбора числа кластеров. При этом предполагается, что значения признаков в исходном наборе данных распределены по нормальному (гауссовскому) закону. Тогда критерием качества кластеризации является близость распределения значений данных внутри кластера к гауссовому, а центры кластеров - к моде распределения.

В процессе работы алгоритма меняется не только расположение центроидов и конфигурации кластеров, но и их число. Алгоритм завершает работу, когда значение статистики критерия соответствия распределения нормальному в кластере окажется выше, чем заданный уровень зна-

чимости (то есть распределение значений в кластере соответствует нормальному в достаточной степени).

Таким образом, выбирать режим автоматического определения числа кластеров имеет смысл только в том случае, если известно, что распределение значений исходных данных близко к нормальному.

Фиксированное число кластеров - следует использовать, если существуют априорные сведения о наличии в исходном наборе данных каких-то групп схожих наблюдений. Тогда число кластеров целесообразно выбрать равным ожидаемому числу таких групп. При этом следует помнить, что при слишком большом (более 5 - 7) числе кластеров модель становится сложной для понимания, хотя, возможно, и более точной.

5. Запуск процесса кластеризации. На данном шаге запускается процесс работы алгоритма кластеризации. Его можно контролировать с помощью ошибок на обучающем и тестовом множествах (рис. 6). Средняя ошибка - это среднее расстояние наблюдений от центров своих кластеров. Очевидно, что чем это расстояние меньше, тем более «тесно» сгруппированы объекты вокруг центров кластеров и тем лучше качество кластеризации (кластеры более выражены и различимы).

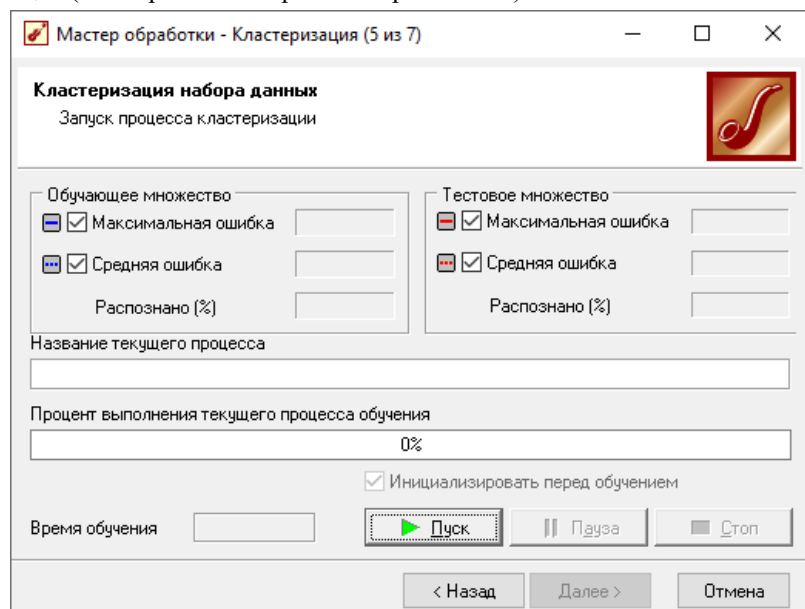


Рис. 6. Запуск процесса обучения

Для запуска процесса обучения следует нажать кнопку «Пуск». В любой момент процесс обучения может быть поставлен на паузу (времен-

но приостановлен). Щелчок по кнопке «Пуск» после кнопки «Пауза» позволит продолжить построение кластерной модели. Кнопка «Стоп» прекращает процесс окончательно, без возможности его продолжения.

6. Определение способов отображения - требуется выбрать способы отображения результатов построения кластерной модели, которые позволяют оценить качество кластеризации и произвести содержательную интерпретацию кластеров (рис. 7).

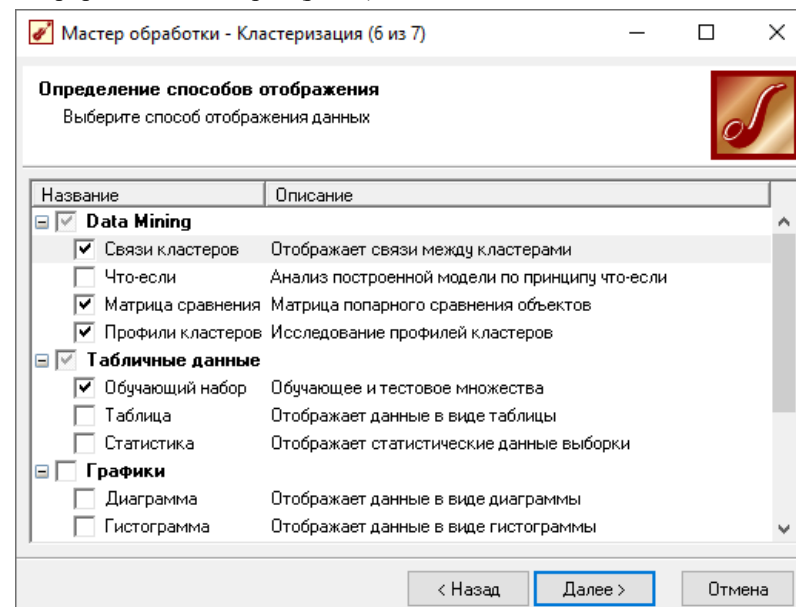


Рис. 7. Выбор способов представления результатов кластеризации

Наиболее подходящими для этих целей являются визуализаторы «Связи кластеров», «Матрица сравнения», «Профили кластеров» и «Обучающий набор». Рассмотрим их более подробно.

Визуализатор «Связи кластеров» (рис. 8)

Визуализатор «Связи кластеров» наглядно отображает кластерную структуру, полученную в результате работы алгоритма k-средних, и позволяет выполнять содержательную интерпретацию кластеров, а также оценить качество кластеризации. На диаграмме каждый кластер представлен прямоугольником, в котором выводятся номер кластера и количество попавших в него наблюдений (мощность кластера).

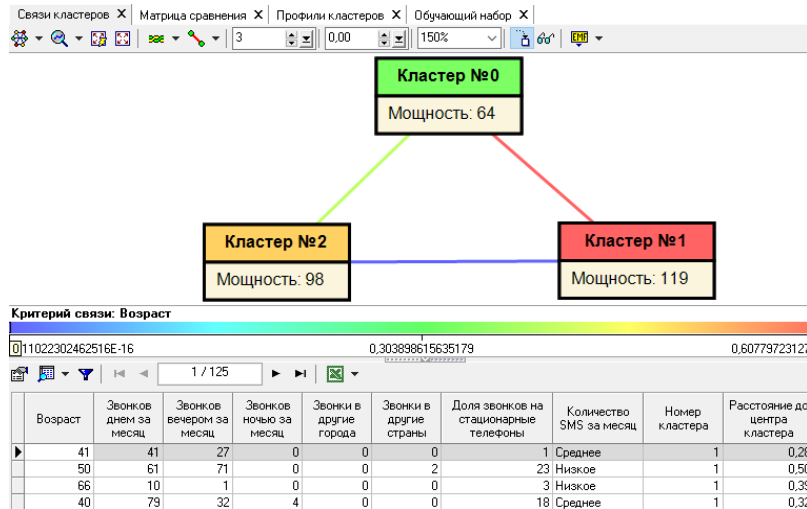



Рис. 8. Визуализатор «Связи кластеров»

Линии между кластерами представляют наличие связи между кластерами, а цвет линии - силу этой связи, которая меняется от 0 до 1. Под диаграммой показана соответствующая цветовая шкала. Чем сильнее связь, тем слабее объекты в двух кластерах отличаются друг от друга, то есть тем менее выражена кластерная структура набора данных. Напротив, чем слабее связь, тем кластеры сильнее отличаются друг от друга. Интуитивно понятно, что чем слабее связи между кластерами, тем лучше работает кластерная модель, построенная с помощью алгоритма k-средних.

Сила связи между кластерами обычно отличается для различных признаков. То есть по одним признакам кластеры могут быть связаны сильнее, а по другим слабее. Чтобы посмотреть силу связи для отдельных признаков, можно воспользоваться кнопкой  - «Критерий связи». В результате откроется список признаков исходного набора данных, из которого следует выбрать тот, связи по которому требуется рассмотреть.

Следует отметить, что если по какому-либо признаку все связи между кластерами сильные (больше 0,7), то можно сделать вывод, что данный признак даёт низкий вклад в кластеризацию и его можно исключить из кластерной модели без существенного ущерба для её точности. Таким образом, диаграмму связей можно использовать для отбора наиболее значимых признаков в кластерную модель. Исключение признаков, по которым кластеры плохо различаются, снижает вычислительные затраты и повышает интерпретируемость модели. На практике можно задать по-

роговое значение силы связи (например, 0,6), исключив из модели все признаки, для которых сила связи превышает этот порог.

Кроме этого, диаграмма связей между кластерами позволяет снизить неопределённость выбора числа кластеров в алгоритме k-средних. Действительно, при изменении числа кластеров будет меняться и конфигурация кластерной структуры, а следовательно, и сила связи между кластерами. Поэтому можно предположить, что с формальной точки зрения лучшим будет то число кластеров, при котором сила связей между ними будет наименьшей. Однако выбранное таким образом число кластеров может противоречить бизнес-логике решаемой задачи.

Визуализатор «Матрица сравнения»

Матрица сравнения (рис. 9) является ещё одним графическим способом отображения силы связи между кластерами.

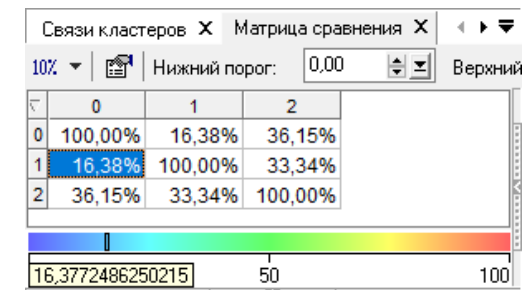


Рис. 9. Визуализатор «Матрица сравнения»

Матрица сравнения представляет собой таблицу, в которой число строк и столбцов равно числу кластеров. Тогда каждая ячейка соответствует паре кластеров, на пересечении строки и столбца которых она расположена. Например, ячейка на пересечении столбца первой строки и второго столбца связывает кластеры с номерами 1 и 2. Исключение составляют ячейки, расположенные по главной диагонали, - они связывают каждый кластер сам с собой.

Тогда в ячейке, связывающей пару кластеров, можно указать степень связи между ними и выделить её соответствующим цветом. Поскольку ячейки на главной диагонали показывают связь кластера с самим собой, то в них будет всегда стоять 100 %.

Визуализатор «Профили кластеров»

Визуализатор «Профили кластеров» представляет собой таблицу (рис. 10), строки которой - поля (признаки) исходного набора данных, а столбцы соответствуют кластерам. В ячейке, расположенной на пересечении строки определённого признака и столбца кластера, указывается показатель, отражающий значимость вклада данного признака в формирование кластера.

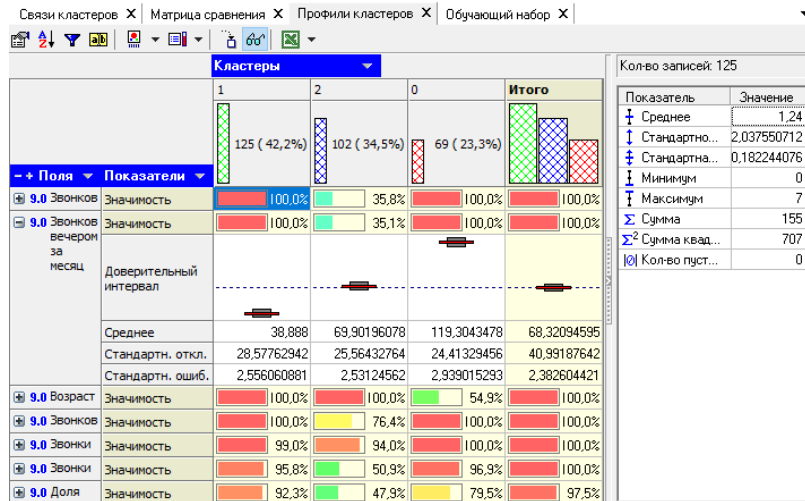


Рис. 10. Визуализатор «Профили кластеров»

Кроме этого, в правой части окна визуализатора отображаются статистические характеристики значений признаков для наблюдений, попавших в определённый кластер. Чтобы увидеть статистику по данному признаку в данном кластере, достаточно выделить соответствующую ячейку, и статистики автоматически актуализируются.

Статистические характеристики очень удобны для оценки выраженности кластерной структуры, полученной с помощью алгоритма k -средних. Например, о различимости кластеров можно судить по степени отличия средних значений признака, для наблюдений попавших в тот или иной кластер.

Визуализатор «Обучающий набор»

Визуализатор «Обучающий набор» представляет собой таблицу, в которой представлены поля исходного набора данных, а также два новых поля, отражающих результаты кластеризации – «Номер кластера» и «Расстояние до центра кластера». Визуализатор «Обучающий набор» удобен в том случае, когда требуется изучить не кластерную структуру в целом, а распределение по кластерам отдельных наблюдений. Для каждого наблюдения здесь указывается номер кластера, в который оно было распределено, а также расстояние данного наблюдения от центра кластера (рис. 11).

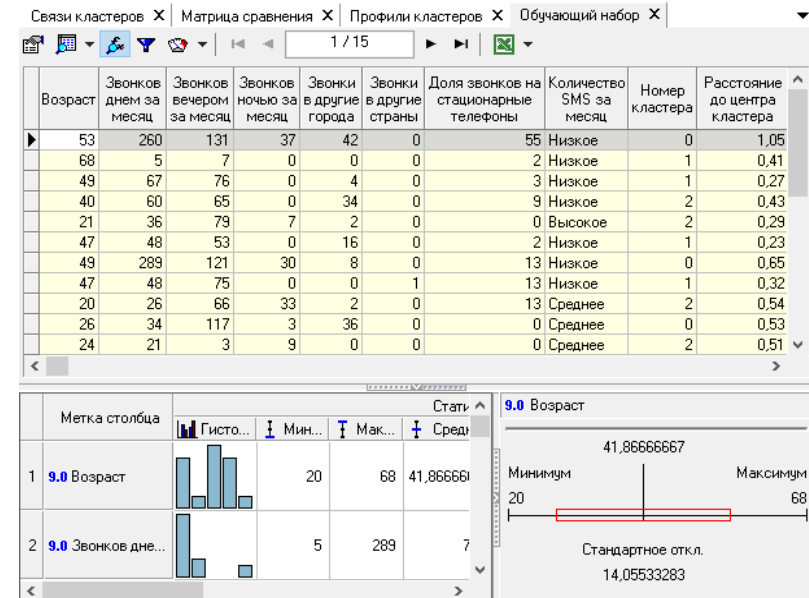


Рис. 11. Визуализатор «Обучающий набор»

Расстояние от наблюдения до центра кластера можно интерпретировать как ошибку. Чем больше это расстояние, тем менее «типичным» объект является для данного кластера и более «похож» на объекты соседнего кластера. При содержательной интерпретации кластеров обобщать свойства кластера на новые объекты, удалённые от его центра, нужно с осторожностью.

С помощью кнопки можно показать (скрыть) отображение статистических характеристик для каждого поля исходного набора данных. Чтобы увидеть статистические характеристики по кластеру, нужно отфильтровать наблюдения по нему. Фильтрация включается/выключается с помощью кнопки .

Порядок выполнения лабораторной работы

1. Загрузить исходный набор данных по указанию преподавателя или в соответствии со своим вариантом.
2. Выполнить пункты 2 - 6, выбрав на шаге 4 опцию «Автоматически определять число кластеров». Повторить, увеличивая уровень значимости, до тех пор, пока алгоритм не начнёт формировать кластеры. Продолжить увеличение уровня значимости на 10 с интервалом 1 и в каждом случае фиксировать сформированное число кластеров. Результаты свести в табл. 1.

Таблица 1

Зависимость числа кластеров от уровня значимости

Уровень значимости									
Число кластеров									

Построить соответствующий график. Сделать выводы о влиянии уровня значимости на число кластеров.

3. Выполнить пункты 2 - 6, выбрав на шаге 4 опцию «Фиксированное количество кластеров». Повторить, изменяя число кластеров от 3 до 7 (5 измерений). Для каждого числа кластеров зафиксировать таблицу сравнения. Сделать вывод о влиянии числа кластеров на силу связи между ними.

4. Выполнить пункты 2 - 6, выбрав на шаге 4 опцию «Фиксированное количество кластеров», и установить число кластеров, равное 3. Зафиксировать силу всех трёх связей для каждого признака. Результаты свести в табл. 2:

Таблица 2
Связи кластеров

Номера кластеров	0-1	0-2	2-1
Признак 1			
Признак 2			
....			

Сделать вывод о том, какие признаки обеспечивают наибольшую и наименьшую степень связи между кластерами.

Контрольные вопросы

1. Опишите общую постановку задачи кластеризации.
2. Приведите примеры использования кластеризации в интеллектуальном анализе данных.
3. Перечислите основные этапы кластерного анализа.
4. Каковы причины популярности алгоритма k-средних?
5. Что такое центроид кластера?
6. Какая величина итеративно минимизируется при работе алгоритма k-средних?
7. В чём заключается инициализация кластеров в алгоритме k-средних?
8. Когда алгоритм k-средних завершает работу?
9. Каковы недостатки алгоритма k-средних?
10. Что отражает уровень значимости? Как он влияет на число кластеров, формируемое алгоритмом?

11. Что можно сказать о признаке, для которого степень связи между кластерами высокая/низкая?

12. Какую информацию содержит матрица сравнения?

13. Что можно сказать о наблюдении, для которого расстояние до центра кластера близко к 1?

Библиографический список

1. Барсегян А.А. Анализ данных и процессов: учеб. пособие / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров -3-е изд., перераб. и доп. - СПб.: БХВ-Петербург, 2009. - 512 с.: ил.
2. Корячко В.П. Интеллектуальные системы и нечеткая логика: учебник / В.П. Корячко, М.А. Бакулева, В.И. Орешков. - Москва: КУРС, 2017. - 346 с.: ил.
3. Осипова Ю.А., Лавров Д.Н. Применение кластерного анализа методом k-средних для классификации текстов научной направленности // Математические структуры и моделирование. 2017 №3(43). С. 108-121.
4. Паклин Н.Б. Бизнес-аналитика: от данных к знаниям (+ CD): учеб. пособие / Н.Б. Паклин, В.И. Орешков. -2-е изд., испр. - СПб.: Питер, 2013. -704 с.