МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

ИССЛЕДОВАНИЕ МЕТОДОВ И АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ БИОМЕДИЦИНСКИХ ДАННЫХ

Методические указания к лабораторной работе

САМАРА 2012

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

ИССЛЕДОВАНИЕ МЕТОДОВ И АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ БИОМЕДИЦИНСКИХ ДАННЫХ

САМАРА 2012

УДК 519.688

Составитель: В.Н. Конюхов

Исследование методов и алгоритмов кластеризации биомедицинских данных. Метод. указания к лабораторной работе/ Самар. гос. аэрокосм. ун-т; Сост. В.Н. Конюхов, Самара, 2012. 25с.

В методических указаниях изложены сведения об основных методах и алгоритмах кластеризации. Показаны области применения, назначение и ограничения. Приведены характеристики качества алгоритмов. Описано применение алгоритмов кластеризации для анализа биомедицинских данных.

Методические указания предназначены для бакалавров, обучающихся по направления подготовки 201000.62 (Биотехнические системы и технологии) и выполняющих лабораторные работы по дисциплине «Автоматизация обработки биомедицинской информации» на кафедре радиотехники и медицинских диагностических систем.

Печатаются по решению редакционно-издательского совета Самарского государственного аэрокосмического университета им. академика С.П. Королёва

Рецензент: проф. Гречишников В.М.

Цель работы: исследование методов и алгоритмов кластеризации медико-биологических данных, их характеристик, ознакомление с пакетами прикладных программ реализующих эти методы, получение практических навыков использования алгоритмов кластеризации (на примере группировки QRS- комплексов электрокардиосигнала).

1. КРАТКИЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

1.1. ОСНОВНЫЕ ПОЛОЖЕНИЯ.

Кластерный анализ предназначен для разбиения совокупности объектов на однородные группы (кластеры[1]) по каким либо схожим признакам, свойствам или по определенному правилу. Соответственно, **кластером** называется множество объектов, объединенных вместе по какому- либо набору свойств или правилу, а процедуру разбиения исходной выборки объектов на множество подгрупп называют кластеризацией.

Кластеризация близка по смыслу к классификации объектов, когда вновь предъявляемый объект относят к одному из известных классов. Однако между этими процедурами существует четкое отличие: если при классификации объект относится к группе по критерию, который имеет смысл в рамках решаемой задачи (например, в задачах медицинской диагностики можно выделить классы гипертоников и людей с нормальным давлением), то кластеризация выполняется посредством некоторой формальной процедуры группировки.

В первом случае, разделение на группы производится на основе некоторой исходной информации, в данном примере, границ артериального давления, установленных до проведения процедуры классификации, исходя из сути решаемой задачи, из накопленного врачебного опыта. Соответственно, построение алгоритма классификации называют в этом случае «обучением с учителем».

При кластеризации неизвестно изначально ни границы между классами, ни само количество классов. Все, что есть в распоряжении исследователя- набор данных и скрытых в них закономерностей. По этой причине кластеризацию называют также «обучением без учителя».

Основная идея кластерного анализа заключается в том, чтобы выделить компактные группы объектов. Например, в биологии ставиться цель разбить животных на различные виды, чтобы содержательно описать различия между ними. В маркетинге это сегментация конкурентов и потребителей. В менеджменте: разбиение персонала на различные по уровню мотивации группы, классификация поставщиков, выявление схожих производственных ситуаций, при которых возникает брак.

[1] От слова cluster- гроздь, группа, скопление.

В общем, всякий раз, когда необходимо классифицировать большой объем исходной информации к пригодным для дальнейшей обработки группам, кластерный анализ оказывается весьма полезным и эффективным. Кроме того, в отличие от многих других статистических процедур, методы кластерного анализа используются тогда, когда каких-либо априорных гипотез относительно классов, когда данных мало и не выполняются требования нормальности распределений случайных величин и другие требования классических методов статистического анализа.

Цели кластеризации могут быть различными в зависимости от особенностей конкретной прикладной задачи:

• Понять структуру множества объектов Xn, разбив его на группы схожих объектов. Упростить дальнейшую обработку данных и принятия решений, работая с каждым кластером по отдельности (стратегия разделяй и властвуй).

• Сократить объём хранимых данных в случае сверхбольшой выборки Xn, оставив по одному наиболее типичному представителю от каждого кластера.

• Выделить нетипичные объекты, которые не присоединяются ни к одному из кластеров. Эту задачу называют одноклассовой классификацией, обнаружением нетипичности или новизны.

В первом случае число кластеров стараются сделать поменьше. Во втором случае важнее обеспечить высокую степень сходства объектов внутри каждого кластера, а кластеров может быть сколько угодно. В третьем случае наибольший интерес представляют отдельные объекты, не вписывающиеся ни в один из кластеров.

Пример кластеризации приведен на рисунке 1. Исходное множество на рисунке 1а, после проведения кластеризации - на рисунке 1б.

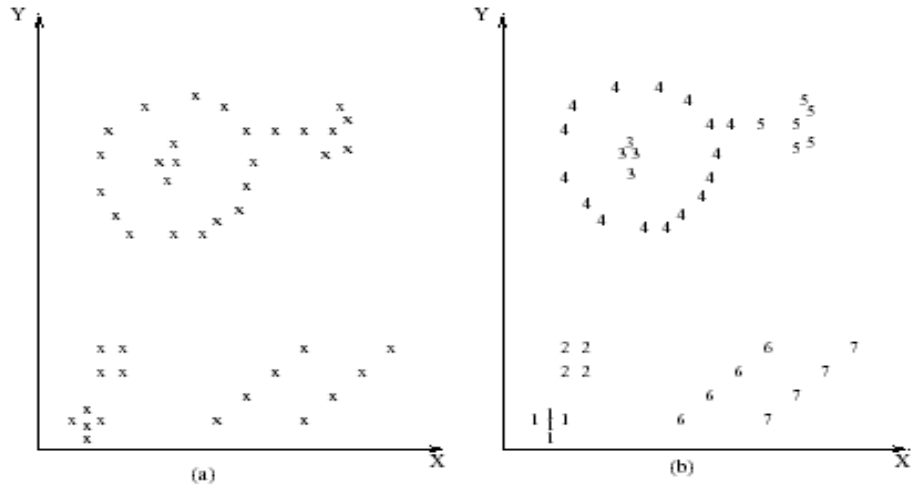

Рисунок 1- Пример кластеризации.

Фактически термин "кластерный анализ" включает в себя достаточно большой набор алгоритмов, используемых при создании классификации. В зависимости от выбранного

алгоритма и его параметров может существенным образом меняться распределение объектов по группам и само количество групп.

1.2. ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ.

Основными элементами, с которыми имеют дело в кластерном анализе, являются **объекты и признаки**. Под объектом подразумевают конкретные предметы исследования, например, пациенты, страдающие теми или иными заболеваниями, микроорганизмы, социальные и экономические системы и т.д. Некоторую совокупность объектов, доступную исследователю для изучения, называют **выборкой**. Количество объектов в такой совокупности принято называть **объемом выборки.**

Признак представляет собой конкретное свойство объекта. Эти свойства могут выражаться как числовыми, так и не числовыми значениями. Например, артериальное давление (систолическое или диастолическое) измеряют в миллиметрах ртутного столба, вес – в килограммах, рост в сантиметрах и т.д. Такие признаки называют количественными признаками. В отличие от этих непрерывных числовых характеристик, ряд признаков может иметь дискретные, прерывистые значения. К ним можно отнести стадии того или иного заболевания, балльные оценки знаний учащихся, 12-балльную шкалу магнитуд землетрясений по Рихтеру, состояние пациента – "здоров" или "болен", пол пациента и т.д. Эти дискретные признаки обычно именуют качественными признаками.

Используя понятия объекта и признака, можно составить прямоугольную таблицу, **матрицу**, состоящую из значений признаков, описывающих свойства исследуемой выборки наблюдений. В данном контексте одно наблюдение будет записываться в виде отдельной строки состоящей из значений используемых признаков. Отдельный же признак в такой матрице данных будет представлен столбцом, состоящим из значений этого признака по всем объектам выборки.

Пример такой матрицы приведен в таблице1. В первом столбце матрицы размещены порядковые номера наблюдений, X1-X6 – некоторые количественные переменные, Качественный признак X7,- характер группы пациентов (здоровые – 1, больные до лечения – 2 и больные после лечения – 3)- который можно использовать для сравнения согласованности результатов кластерного анализа с результатами обследования.

Таблица 1

№	*X1*	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*
1	80,93	0,60	0,30	4,94	1,21	5,85	1
2	80,30	0,50	0,70	5,10	1,30	5,70	1
3	80,22	0,56	0,56	5,59	1,09	5,29	2

4	80,80	0,30	0,60	4,90	1,20	5,90	2
5	80,00	0,50	0,90	5,20	1,10	5,80	1
6	80,60	0,70	0,30	5,10	1,20	5,90	2
7	79,50	0,50	0,60	5,30	2,00	5,30	3
8	81,40	0,60	0,50	5,30	1,80	3,90	2
9	80,00	0,30	0,80	4,90	1,80	5,10	3
10	80,50	0,50	0,50	5,10	1,90	4,90	3

Основными вопросами при проведении кластеризации являются два:

1) как установить подобие объектов;
2) как оценить качество кластеризации.

1.3. МЕРЫ ПОДОБИЯ ОБЪЕКТОВ.

Наиболее естественно в качестве меры подобия объектов выбрать расстояние в пространстве признаков объекта. Если расстояние выбрано удачно, то можно ожидать, что его значения между кластерами будут гораздо больше, чем расстояния между объектами внутри кластера.

Расстоянием или **метрикой** между объектами с номерами i и k в пространстве признаков называется такая величина d_{ik}, которая удовлетворяет следующим аксиомам:

1. $d_{ik} \geq 0$ (расстояние не отрицательно);

2. $d_{ik} = d_{ki}$ (симметрия);

3. $d_{ik} \leq d_{ij} + d_{jk}$ (неравенство треугольника);

4. Если $d_{ik} \neq 0$, то $i \neq k$ (различимость нетождественных объектов);

5. Если $d_{ik} = 0$, то $i = k$ (неразличимость тождественных объектов)

Меру близости или **степень подобия** объектов удобно представить как обратную величину от расстояния между объектами, что вполне логично, чем меньше расстояние, тем подобнее объекты и, для совпадающих объектов, степень подобия равна бесконечности.

Способы задания расстояния могут быть различными. Существуют около 50 различных метрик, применяемых в кластерном анализе. Наиболее естественным и интуитивно понятным в случае количественных признаков является **евклидово расстояние:**

$$d_{ik} = \left(\sum_{j=1}^{m} \left(x_{ij} - x_{kj} \right)^2 \right)^{\frac{1}{2}},$$

где d_{ik} – расстояние между i-ым и k-ым объектами;

x_{ij} и x_{kj} – численное значение j-ой переменной для i-го и k-го объекта соответственно;

m – количество переменных, которыми описываются объекты.

Вместо обычного евклидового расстояния на практике часто используют его квадрат d_{ik}^2.

Целью использования квадрата евклидова расстояния, является придание больших весов более отдаленным друг от друга объектам. Кроме того, в ряде случаев используется "взвешенное" евклидово расстояние, при вычислении которого для отдельных слагаемых используются весовые коэффициенты, которые определяют значимость признака. Последнее особенно важно при проведении медико-биологических исследований, так как, зачастую, признаки, описывающие биологический объект, имеют различный физический и биологический смысл.

В качестве других метрик, часто применяющихся при проведении кластерного анализа можно упомянуть расстояние городских кварталов (манхэттенское расстояние), расстояние Чебышева, степенное расстояние, процент несогласия.

Расстояние городских кварталов (манхэттенское расстояние). Это расстояние является просто средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного евклидова расстояния. Однако, для этой меры влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат). Манхэттенское расстояние вычисляется по формуле:

$$d_{ik} = \left(\sum_{j=1}^{m} \left| x_{ij} - x_{kj} \right| \right).$$

Расстояние Чебышева. Это расстояние может оказаться полезным, когда желают определить два объекта как "различные", если они различаются по какой-либо одной координате (каким-либо одним измерением). Расстояние Чебышева вычисляется по формуле:

$$d_{ik} = \max \left| x_{ij} - x_{kj} \right| .$$

Степенное расстояние. Иногда желают прогрессивно увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Это может быть достигнуто с использованием степенного расстояния. Степенное расстояние вычисляется по формуле:

$$d_{ik} = \left(\sum_{j=1}^{m} \left(x_{ij} - x_{kj} \right)^{r} \right)^{\frac{1}{p}}.$$

где r и p - параметры, определяемые пользователем. Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, параметр r ответственен за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра - r и p, равны двум, то это расстояние совпадает с евклидовым расстоянием.

Процент несогласия. Эта мера используется в тех случаях, когда данные являются категориальными. Это расстояние вычисляется по формуле:

$$d_{ik} = \left(\sum_{j=1}^{m} \left(x_{ij} \neq x_{kj} \right) \right) \cdot 1/j .$$

Другой важной величиной в кластерном анализе является расстояние между целыми группами объектов. Пример различного способа задания такого расстояния приведены на рисунке 2.

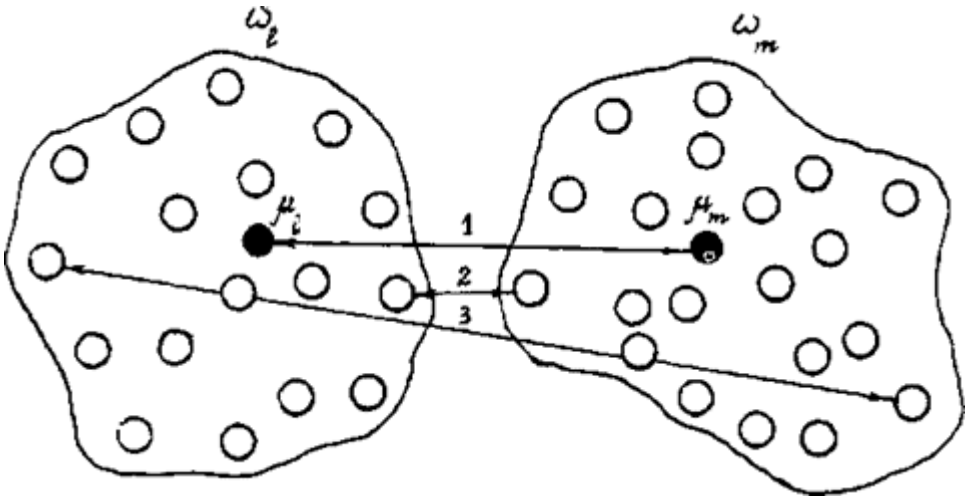

Рисунок 2- Различные способы определения расстояния между кластерами wl и wm: 1 — по центрам тяжести, 2 — по ближайшим объектам, 3 — по самым далеким объектам.

Расстояние ближайшего соседа есть расстояние между ближайшими объектами кластеров. Расстояние дальнего соседа – это расстояние между самыми дальними объектами кластеров. Расстояние центров тяжести равно расстоянию между центральными точками кластеров, опредяляемыми как математическое ожидание.

Выбор той или иной меры расстояния между кластерами влияет, главным образом, на вид выделяемых алгоритмами кластерного анализа геометрических группировок объектов в пространстве признаков. Так, алгоритмы, основанные на расстоянии ближайшего соседа, хорошо работают в случае группировок, имеющих сложную, в частности, цепочечную структуру. Расстояние дальнего соседа применяется, когда искомые группировки образуют в пространстве признаков шаровидные облака. И промежуточное место занимают алгоритмы, использующие расстояния центров тяжести и средней связи, которые лучше всего работают в случае группировок эллипсоидной формы.

1.4. КРИТЕРИИ КАЧЕСТВА ГРУППИРОВКИ.

Не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд эвристических критериев, а также ряд алгоритмов, не имеющих чётко выраженного

критерия, но осуществляющих достаточно разумную кластеризацию «по построению». Все они могут давать разные результаты.

Критерий качества кластеризации в той или иной мере отражает следующие неформальные требования:

а) внутри групп объекты должны быть тесно связаны между собой;

б) объекты разных групп должны быть далеки друг от друга;

в) при прочих равных условиях распределения объектов по группам должны быть равномерными.

Требования а) и б) выражают стандартную концепцию компактности классов разбиения; требование в) состоит в том, чтобы критерий не навязывал объединения отдельных групп объектов.

Простейший критерий качества непосредственно базируется на величине расстояния между кластерами. Однако такой критерий не учитывает "населенность" кластеров - относительную плотность распределения объектов внутри выделяемых группировок. Поэтому другие критерии основываются на вычислении средних расстояний между объектами внутри кластеров. Но наиболее часто применяются критерии в виде отношений показателей "населенности" кластеров к расстоянию между ними. Это, например, может быть отношение суммы межклассовых расстояний к сумме внутриклассовых (между объектами) расстояний или отношение общей дисперсии данных к сумме внутриклассовых дисперсий и дисперсии центров кластеров.

Функционалы качества и конкретные алгоритмы автоматической классификации достаточно полно и подробно рассмотрены в специальной литературе. Эти функционалы и алгоритмы характеризуются различной трудоемкостью и подчас требуют ресурсов высокопроизводительных компьютеров.

2. МЕТОДЫ И АЛГОРИТМЫ КЛАСТЕРНОГО АНАЛИЗА.

2.1. КЛАССИФИКАЦИЯ МЕТОДОВ И АЛГОРИТМОВ КЛАСТЕРНОГО АНАЛИЗА.

Алгоритмы и методы кластерного анализа отличаются большим разнообразием. Это могут быть, например, алгоритмы, реализующие полный перебор сочетаний объектов или осуществляющие случайные разбиения множества объектов. В то же время большинство таких алгоритмов состоит из двух этапов. На первом этапе задается начальное (возможно, искусственное или даже произвольное) разбиение множества объектов на классы и определяется некоторый математический критерий качества автоматической классификации. Затем, на втором этапе, объекты переносятся из класса в класс до тех пор, пока значение критерия не перестанет улучшаться.

Основная проблема выбора конкретного метода и какого-либо алгоритма его реализации заключается в том, что трудно описать кластеры, которые ожидается получить в результате работы соответствующих программ.

Все методы можно условно разделить на три типа:

1) эвристический: задается точное определение требуемого "образа" кластеров (например, однородные точки должны находиться внутри гиперсферы радиуса R и некоего "центра тяжести"). К недостаткам этого подхода относится то, что определение типичного образа однородных групп может оказаться слишком строгим, т. е. принципиально допустимые кластеры, вид которых не соответствует критерию, отклоняются;

2) оптимизационный: требуемое разбиение соответствует минимуму заданного функционала качества. Этот подход интересен тем, что он дает чисто математическую постановку задачи классификации. Но тогда проблема заключается в выборе и выражении функционала, что чаще всего оказывается нетривиальным (не говоря о самом процессе нахождения экстремума);

3) аппроксимационный: отыскивается такое преобразование (представление) множества данных, которое раскрывает его структуру как состоящую из отдельных областей.

Эти три типа кластеризации связаны тем, что они все носят оптимизационный характер, ведь можно сформулировать первый и третий подходы (а, в общем, любой подход: речь идет о математическом формализме) в виде минимизации некоего функционала.

В зависимости от выбранного признака алгоритмы кластеризации можно классифицировать следующим образом.

<u>Классификация по результатам процедуры:</u>

1) разбиение с непересекающимися классами. Результат представлен в виде кластеров: все объекты внутри найденного класса считаются тождественными, а объекты разных классов – нет;

2) разбиение с пересекающимися классами. В этом случае результаты классификации заданы:

a) либо введением степени принадлежности объекта к классу в духе теории нечетких множеств;

b) либо определением вероятности принадлежности объекта к классу;

c) либо простым перечнем объектов в зоне пересечения.

Однако такой результат может считаться неудовлетворительным для целого ряда применений, где требуется четкое разбиение множества на разные классы;

3) иерархическое дерево. Оно указывает, на какой ступени можно объединить объекты друг с другом, в зависимости от "глубины" узла объединения в дереве. На минимальной – все

объекты отдельны; на максимальной – все охваченными в одном классе. Этот подход не полностью решает задачу кластеризации, поскольку после его завершения остается применить критерии, чтобы обрисовать разные классы.

Классификация по степени участия человека в процессе кластеризации:

1) человек не принимает участия в работе алгоритма: классификация производится чисто машинным способом. Тем не менее, исследователю принадлежит право отобрать меру расстояния и параметры классификации;
2) человек участвует в процессе: он принимает решения о разбиении на основании информации о классификации, которая выдана алгоритмом. При этом нужно эффективное визуальное представление данных. Одним из этих методов является метод упорядочения матрицы связи.

Классификация по заданным условиям:

1) классификация свободная: нет априорных сведений;
2) классификация с заданным числом кластеров;
3) классификация с заданным порогом внутригруппового сходства.

Классификация по структуре построения алгоритмов:

1) иерархические;
2) неиерархические.

2.2. ИЕРАРХИЧЕСКИЕ МЕТОДЫ КЛАСТЕРИЗАЦИИ.

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие кластеры или разделении больших кластеров на меньшие кластеры. Различают два типа иерархической кластеризации- сверху вниз (делимые методы, divisive anaysis) или снизу вверх (агломеративные, agglomerative nesting).

В первом случае в начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Во втором- все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер. Схематически принцип работы этих разновидностей методов может быть представлен следующим образом (рис. 3).

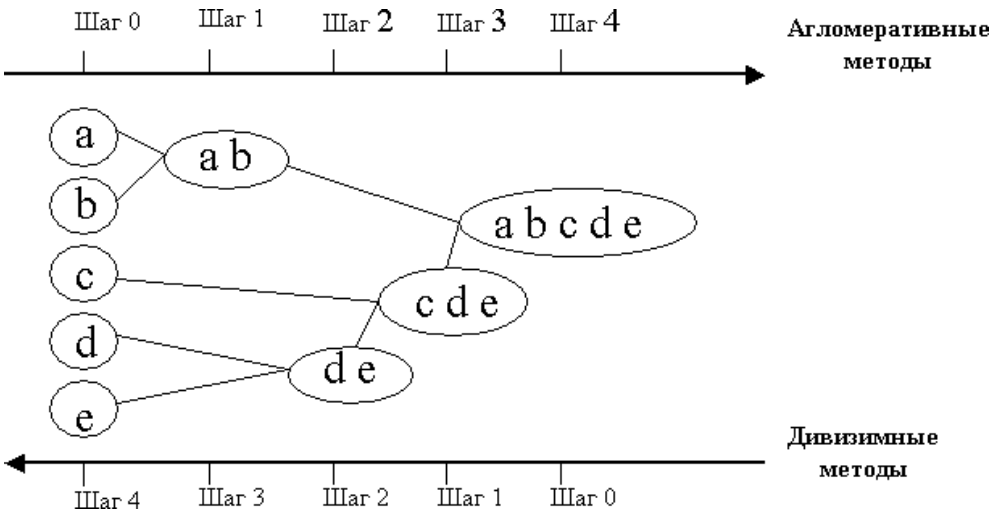

Рисунок 3- Схема иерархических методов кластеризации.

Иерархические методы кластеризации различаются правилами построения кластеров. В качестве правил выступают критерии, которые используются при решении вопроса о "схожести" объектов при их объединении в группу (агломеративные методы) либо разделения на группы (дивизимные методы).

Иерархические методы кластерного анализа используются при небольших объемах наборов данных. Преимуществом иерархических методов кластеризации является их наглядность.

Иерархические алгоритмы связаны с построением дендрограмм (от греческого dendron - "дерево"), которые являются результатом иерархического кластерного анализа. Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.

Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры.

Например, в таблице 1 содержится 10 наблюдений и, соответственно, получается 100 значений расстояний. После вычисления матрицы расстояний начинается процесс агломерации, проходящий последовательно шаг за шагом. На первом шаге этого процесса два исходных наблюдения, между которыми самое минимальное расстояние, объединяются в один кластер, состоящий уже из двух объектов. Таким образом, вместо бывших n монокластеров (кластеров, состоящих из одного объекта) после первого шага останется $(n-1)$ кластеров, из которых один кластер будет содержать два объекта, а $(n-2)$ кластеров будут по-прежнему состоять всего лишь из одного объекта. Отметим, что на втором шаге возможны различные методы объединения между собой $(n-2)$ кластеров. Это вызвано тем, что один из этих кластеров уже содержит два объекта. По этой причине возникает два основных вопроса:

- как вычислять координаты кластера из двух (а далее и более двух) объектов,

- как вычислять расстояние до таких "полиобъектных" кластеров от "монокластеров" и между "полиобъектными" кластерами?

Эти вопросы, в конечном счете, и определяют окончательную структуру итоговых кластеров (под структурой кластеров подразумевается состав отдельных кластеров и их взаимное расположение в многомерном пространстве).

На втором шаге, в зависимости от выбранных методов вычисления координат кластера состоящего из нескольких объектов и способа вычисления межкластерных расстояний, возможно либо повторное объединение двух отдельных наблюдений в новый кластер, либо присоединение одного нового наблюдения к кластеру, состоящему из двух объектов. Для удобства большинство программ агломеративно-иерархических методов по окончании работы могут предоставить для просмотра графики (см. рисунок 4). Эти графики отражают процесс агломерации, слияния отдельных наблюдений в единый окончательный кластер.

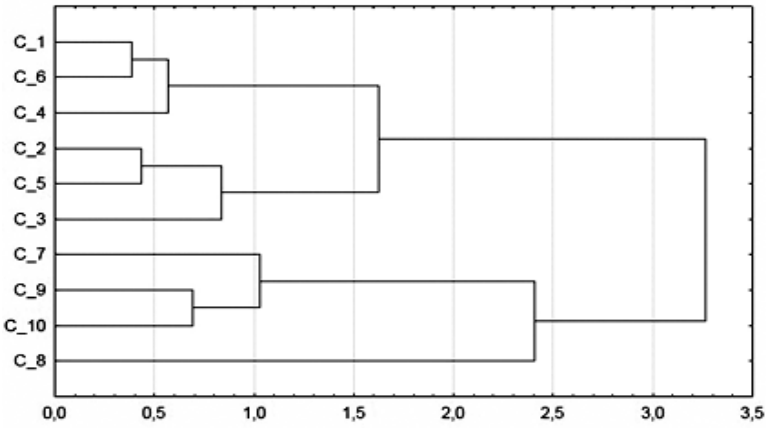

Рисунок 4- Результат иерархической кластеризации данных таблицы 1. Горизонтальная ось представляет собой расстояние (например, евклидово).

2.3. НЕИЕРАРХИЧЕСКИЕ МЕТОДЫ КЛАСТЕРИЗАЦИИ.

При большом количестве наблюдений иерархические методы кластерного анализа не пригодны. В таких случаях используют неиерархические методы, основанные на разделении, которые представляют собой итеративные методы дробления исходной совокупности. В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки.

Такая неиерархическая кластеризация состоит в разделении набора данных на определенное количество отдельных кластеров. Существует два подхода. Первый заключается в определении границ кластеров как наиболее плотных участков в многомерном пространстве исходных данных, т.е. определение кластера там, где имеется большое "сгущение точек". Второй подход заключается в минимизации меры различия объектов.

Наиболее распространен среди неиерархических методов алгоритм k-средних, также

называемый быстрым кластерным анализом. В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для возможности использования этого метода необходимо иметь гипотезу о наиболее вероятном количестве кластеров.

Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.

Работу алгоритма можно разбить на два этапа:

1) первоначальное распределение объектов по кластерам;

2) итеративный процесс.

На первом этапе выбирается число k, и на первом шаге эти точки считаются "центрами" кластеров. Каждому кластеру соответствует один центр. Выбор начальных центров может осуществляться следующим образом:

1) выбор k-наблюдений для максимизации начального расстояния;

2) случайный выбор k-наблюдений;

3) выбор первых k-наблюдений.

В результате каждый объект назначен определенному кластеру.

На втором этапе вычисляются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются. Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:

1) кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации;

2) число итераций равно максимальному числу итераций.

Пример работы алгоритма приведен на рисунке 5.

Достоинства алгоритма k-средних:

-простота использования;

-быстрота использования;

-понятность и прозрачность алгоритма.

Недостатки алгоритма k-средних:

-алгоритм слишком чувствителен к выбросам, которые могут искажать среднее. Возможным решением этой проблемы является использование модификации алгоритма - алгоритм k-

медианы;

-алгоритм может медленно работать на больших базах данных. Возможным решением данной проблемы является использование выборки данных.

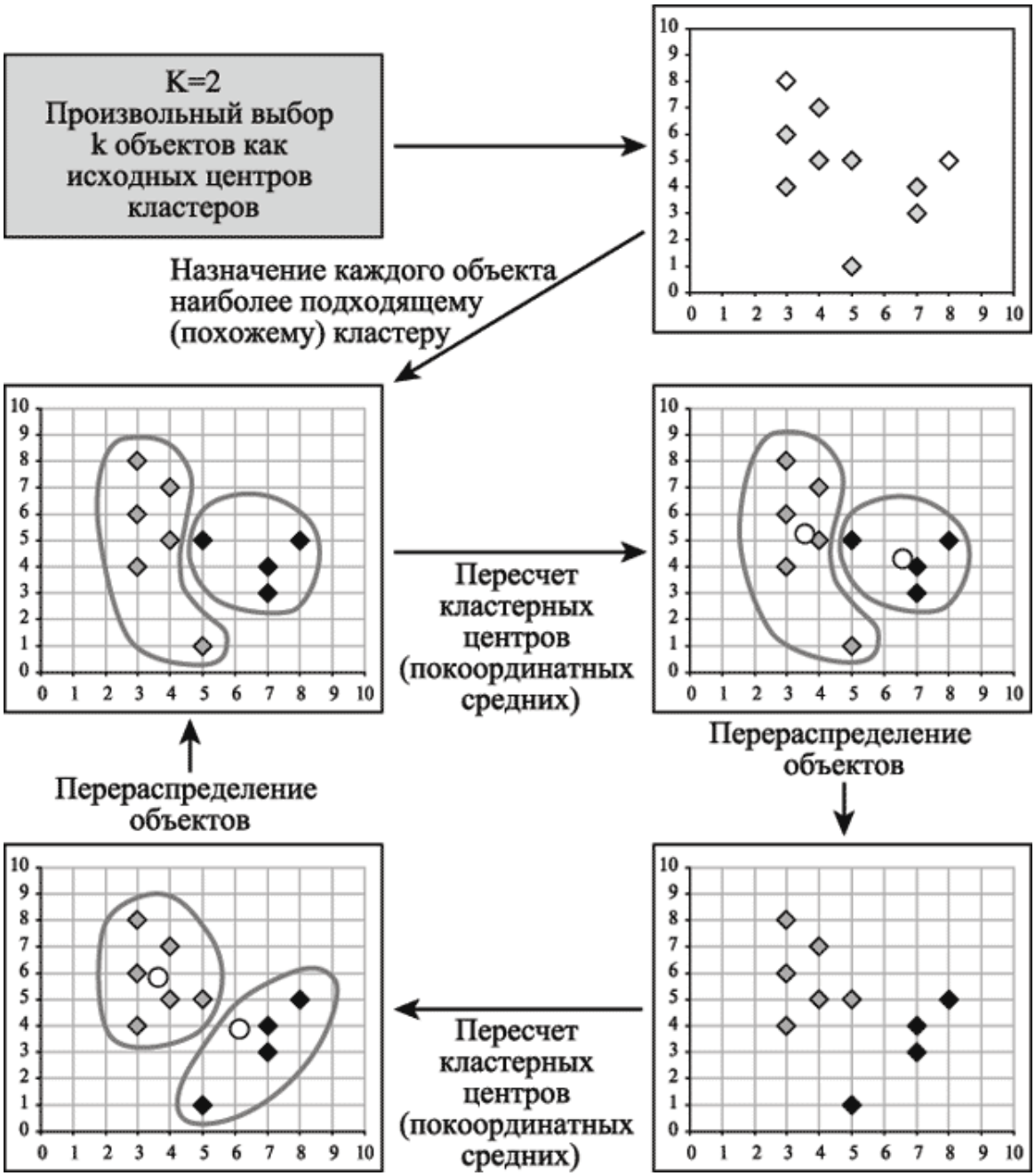

Рисунок 5- Пример работы алгоритма к- средних.

2.4. АЛГОРИТМ ПОРОГОВОГО РАССТОЯНИЯ.

Алгоритм порогового расстояния относится к неиерархическим методам. Алгоритм работает следующим образом:

1) задаем некоторое пороговое расстояние;

2) выбираем произвольный объект и назначаем его центром кластера;

3) рассчитываем расстояние от первого выбранного объекта до второго. Если расстояние меньше порогового, то второй объект включаем в первый кластер. Если расстояние больше порогового, то второй объект задает центр нового кластера;

4) выбираем третий объект и рассчитываем расстояния до центров полученных кластеров. Решение принимаем в соответствии с пунктом 3;

5) повторяем процедуру для всех объектов.

Результаты работы алгоритма существенно зависят от выбора первого объекта и порогового расстояния. При малой величине порогового расстояния каждый объект может сформировать свой кластер и наоборот. Поэтому пороговое расстояние должно быть больше, чем типичное внутрикластерное расстояние и меньше, чем межкластерное. На практике проводят несколько процедур кластеризации с различными значениями порогового расстояния и анализируют результаты.

2.4. НЕКОТОРЫЕ ПРАКТИЧЕСКИЕ ЗАМЕЧАНИЯ ПРИ ПРОВЕДЕНИИ КЛАСТЕРНОГО АНАЛИЗА.

В общем случае, процедура кластеризации включает в себя четыре этапа (рис.6):

1. выделение характеристик объекта;
2. определение метрики;
3. разбиение объектов на группы;
4. представление результатов.

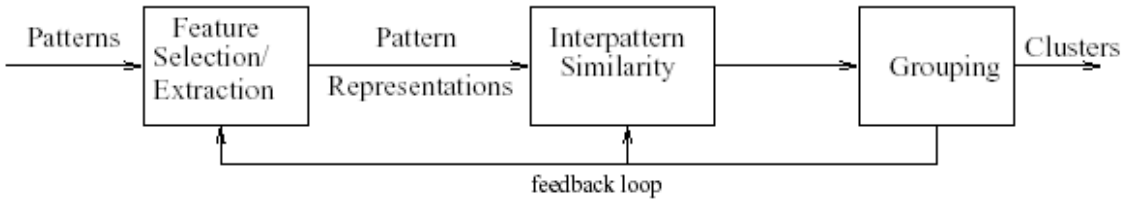

Рисунок 6- Общая схема кластеризации.

Критически важным этапом для получения осмысленных результатов кластеризации является этап выделения характеристик объекта (feature extraction). При большом числе переменных, описывающих объект, появляются кластеры нечеткой структуры. В результате достаточно сложно интерпретировать полученные кластеры. Кроме того, с ростом размерности пространства признаков, резко возрастают вычислительные затраты.

Более понятные и прозрачные результаты кластеризации могут быть получены, если вместо множества исходных переменных использовать некие обобщенные переменные или критерии, содержащие в сжатом виде информацию о связях между переменными. Т.е. возникает задача понижения размерности данных. Решение этой задачи может проводиться

различными методами, например, методами факторного анализа. Возможно также использование эвристических методик понижения размерности.

Рассмотрим следующий пример. Предположим, что есть входные вектора отсчетов QRS- комплекса электрокардиосигнала, полученные из суточной записи электрокардиосигнала, размерностью n=64 для каждого вектора, где n- количество отсчетов сигнала. Допустим, наша задача заключается в определении количества различных типов QRS- комплексов в этой записи. Можно представить каждый QRS- комплекс точкой в 64-мерном пространстве, рассчитать расстояния между комплексами и определить по одному из выбранных алгоритмов кластеризации разбиение этого пространства. Тогда, если в среднем в суточной записи около 100000 QRS- комплексов, то, при условии выбора в качестве меры евклидова расстояния, необходимо подсчитать 6400000 разностей, возвести их в квадрат, просуммировать и извлечь квадратный корень.

С другой стороны, если предварительно взять в качестве признака QRS- комплекса значения его длительности, амплитуды и полярности, то пространство признаков буде равно трем. Соответственно, существенно уменьшиться вычислительная сложность.

3. ПАКЕТЫ ПРОГРАММ ДЛЯ ПРОВЕДЕНИЯ КЛАСТЕРНОГО АНАЛИЗА.

Практически в любом современном математическом пакете программ, таком как STATISTICA, STADIA, SPSS, MATLAB и др., присутствуют процедуры для проведения кластерного анализа. Так, в пакете MATLAB, определены следующие функции для кластерного анализа:

cluster - деление иерархического дерева кластеров (группировка выходных данных функции linkage) на отдельные кластеры;

clusterdata - группировка матрицы исходных данных в кластеры;

cophenet - расчет коэффициента качества разбиения исходных данных на кластеры (этот коэффициент можно рассматривать как аналог коэффициента корреляции, чем его значение ближе к 1, тем лучше выполнено разбиение на кластеры);

dendrogram - дендрограмма кластеров;

inconsistent - расчет коэффициентов несовместимости для каждой связи в иерархическом дереве кластеров и может использоваться как оценка качества разбиения на кластеры;

kmeans - кластеризация на основе внутригрупповых средних;

linkage - формирование иерархического дерева бинарных кластеров;

pdist - расчет парных расстояний между объектами (векторами) в исходном множестве данных;

silhouette - график силуэта кластеров;

squareform - преобразование вектора выходных данных функции pdist в симметричную

квадратную матрицу.

Подробную информацию об этих функциях можно получить, воспользовавшись справкой Matlab или на сайте http://matlab.exponenta.ru/statist/book2/index.php. Так, например, функция clusterdata(X, CUTOFF), производит иерархическую кластеризацию данных X, представленных матрицей MxN, где M-число наблюдений N переменных (см. приложение А). Переменная CUTOFF фактически задает количество кластеров.

4. ОПИСАНИЕ ДАННЫХ И ЭКСПОРТ ДАННЫХ В ПАКЕТ MATLAB

Данные представляют собой записи нормальных QRS- комплексов и желудочковых экстрасистол в формате Matlab. Длина каждой записи составляет 64 отсчета. Соответственно, каждый QRS- комплекс можно представить точкой в 64-х мерном пространстве. Записи желудочковых экстрасистол находятся в файлах 1V.mat-20V.mat, записи нормальных QRS-комплексов- в файлах 1.mat-20.mat. На рисунке 7 приведен нормальный QRS- комплекс, на рисунке 8- желудочковая экстрасистола.

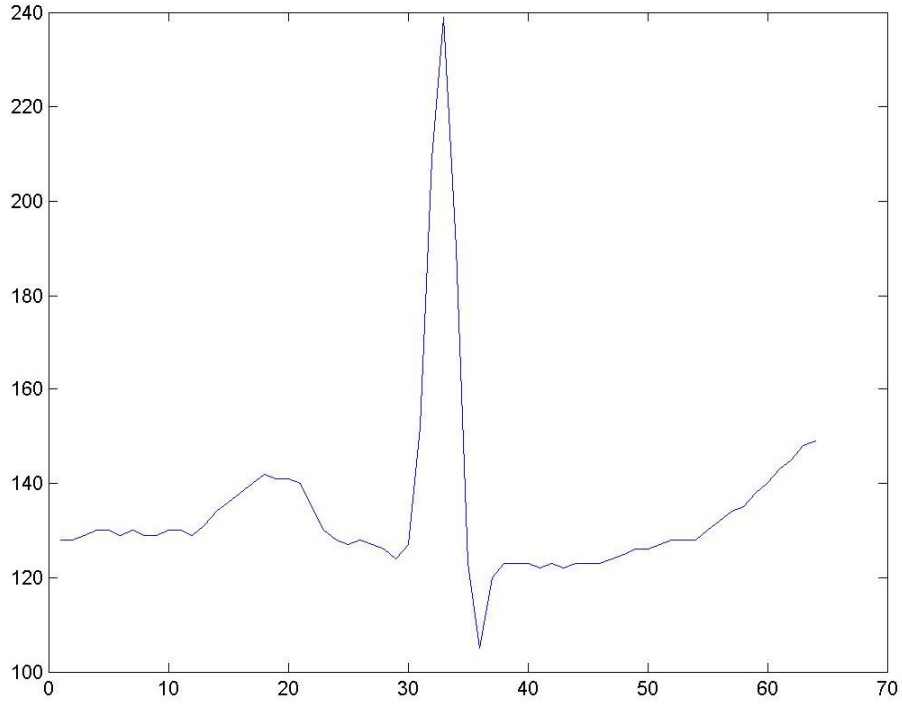

Рисунок 7- Нормальный QRS- комплекс

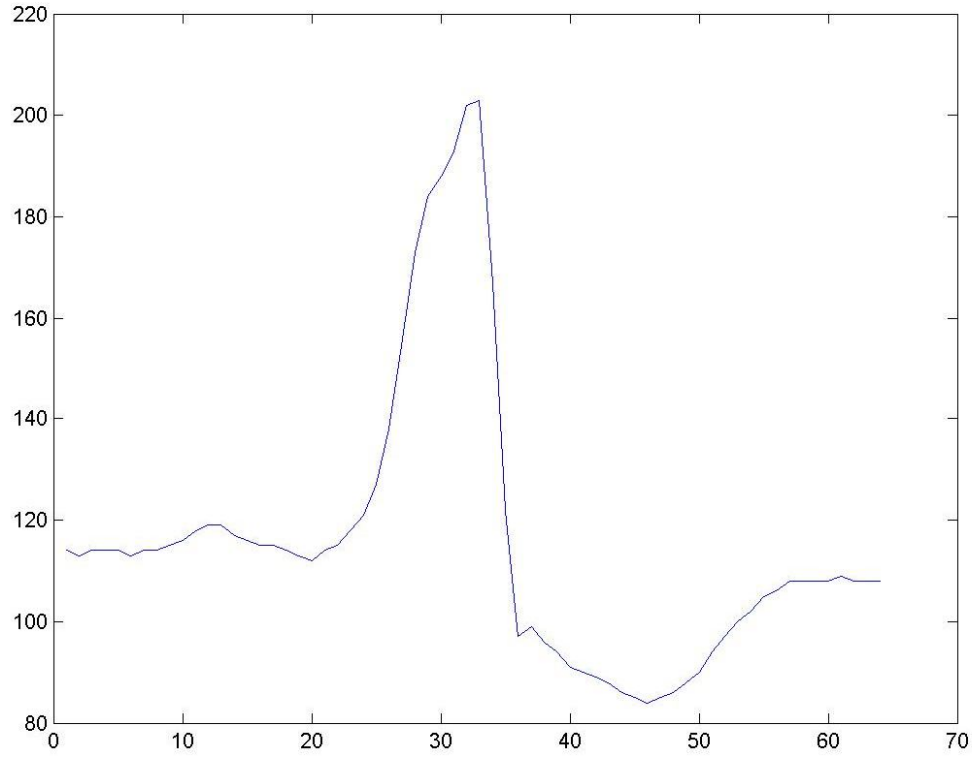

Рисунок 8- Желудочковая экстрасистола.

Загрузить файлы в среду Matlab можно с помощью команды load. Например, load 1 загружает данные из файла 1.mat и автоматически создает переменную в рабочем пространстве (workspace) A1. Соответственно load 1V загружает файл 1V.mat и создает переменную A1v. Просмотреть загруженные данные можно с помощью команды plot. Так, plot (A1v) выведет желудочковую экстрасистолу, хранящуюся в файле 1V.mat.

Визуально видно, что существуют различия между QRS- комплексами на рисунках 7 и 8 и они могут быть отнесены к разным группам (например, длительность желудочковой экстрасистолы примерно в 2 раза больше нормального комплекса).

В данной лабораторной работе, в предположении, что неизвестна классификация комплексов, необходимо разбить исходные файлы с помощью алгоритма кластеризации на группы и оценить качество такого разбиения.

В качестве исходного пространства признаков использовать:

1) вектор исходного сигнала (размерность пространства=64);

2) описании QRS-комплекса в пространстве меньшей размерности. В качестве параметров сжатого описания выбрать максимальное значение, дисперсию и среднее значение QRS-комплекса (размерность пространства=3).

5. ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ

1. Загрузите в рабочее пространство Matlab данные из файлов 1.mat-10.mat и 1V.mat-10V.mat.

2. Сформируйте из загруженных данных матрицу MxN, где M- номер QRS-комплекса, N- отсчеты QRS-комплекса (см. приложение В).

3. Преобразуйте матрицу с помощью команды rot90 (см. приложение В).

4. С помощью команды **clusterdata** определите число кластеров и количество объектов в каждом кластере для значений переменной **cutoff=** 2, 3, 4, 5, 10, 15, 20.

5. С помощью последовательности команд Y = pdist(A,'euclid'); Z = linkage(Y, 'single'); dendrogram(Z) постройте дендрограмму для матрицы, полученной в п.3.

6. Для каждого из файлов 1.mat-10.mat и 1V.mat-10V.mat определить максимальное значение, среднеквадратичное отклонение и среднее значение (использовать команды max, std и mean).

7. Построить матрицу MxN, где M- номер QRS-комплекса, N- параметры сжатого описания (максимальное значение, среднеквадратичное отклонение и среднее значение).

8. С помощью команды **clusterdata** определите число кластеров и количество объектов в каждом кластере для значений переменной **cutoff=** 2, 3, 4, 5, 10, 15, 20.

9. С помощью последовательности команд Y = pdist(A,'euclid'); Z = linkage(Y, 'single'); dendrogram(Z) постройте дендрограмму для матрицы, полученной в п.7.

10. Постройте на одном графике параметры QRS-комплексов для файлов 1.mat-10.mat и 1V.mat-10V.mat. По оси абсцисс отложить среднее значение, по оси ординат-среднеквадратичное отклонение.

6. СОДЕРЖАНИЕ ОТЧЕТА

1. Наименование и цель работы.
2. Матрица, полученная в п.3.
3. Результаты п.4.
4. Матрица, полученная в п.6.
5. Дендрограммы п.5 и п.9.
6. Результаты п.7.
7. График параметров QRS-комплексов полученный в п.8.
8. Выводы.

7. КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Для чего предназначен кластерный анализ?
2. Дайте определение кластеру.
3. В чем различие между кластеризацией и классификацией?
4. В каких случаях целесообразно применение кластерного анализа?
5. Какие основные цели кластеризации?

6. Что подразумевается под объектом и признаком при проведении кластеризации? В каких шкалах измеряются признаки?

7. Какие основные вопросы необходимо решить при проведении кластеризации?

8. Что обычно выбирается в качестве меры подобия объектов?

9. Дайте определение расстояния между объектами.

10. Что такое степень подобия объектов?

11. Как определяется евклидово расстояние?

12. В каких случаях полезно использовать расстояние Чебышева?

13. Приведите выражение для степенного расстояния.

14. Какая мера может использоваться в случае категориальных данных?

15. Какими способами можно задать расстояние между группами объектов?

16. На что влияет выбор меры расстояния между кластерами?

17. Какие требования отражают критерии качества кластеризации?

18. На какие основные типы можно разделить методы кластеризации?

19. Как классифицируются алгоритмы кластеризации по результатам процедуры?

20. Как классифицируются алгоритмы кластеризации по степени участия человека?

21. Как классифицируются алгоритмы кластеризации по заданным условиям?

22. Как классифицируются алгоритмы кластеризации по структуре построения?

23. В чем суть иерархической кластеризации?

24. Какие типы иерархической кластеризации существуют? В чем их различие?

25. В каких случаях используются методы иерархической кластеризации? В чем их преимущество?

26. Что такое дендрограмма?

27. В чем суть методов неиерархической кластеризации?

28. Поясните, в чем заключается алгоритм k- средних?

29. Какие достоинства и недостатки имеет алгоритм k- средних?

30. Поясните, в чем заключается алгоритм порогового расстояния?

31. Какие этапы, в общем случае, включает в себя процедура кластеризации?

32. Поясните, для чего необходимо понижение размерности пространства признаков при проведении кластерного анализа?

СПИСОК ЛИТЕРАТУРЫ

1. Половко А. М., Бутусов П. Н. MATLAB для студента. — СПб.: БХВ-Петербург, 2005. - 320 с.
2. О.Ю. Реброва. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. М., МедиаСфера, 2003. 312с.

3. Новиков Д.А., Новочадов В.В. Статистические методы в медико-биологическом эксперименте (типовые случаи). Волгоград: Издательство ВолГМУ, 2005. – 84 с.

ПРИЛОЖЕНИЕ А

Описание

CLUSTERDATA Construct clusters from data.
T = CLUSTERDATA(X, CUTOFF) construct clusters from a given data X.
X is a matrix of size M by N, treated as M observations of N
variables. CUTOFF is a threshold for cutting the hierarchical
tree generated by LINKAGE into clusters. When 0 < CUTOFF < 1,
Clusters are formed when inconsistent values are greater than
CUTOFF (see INCONSISTENT). When CUTOFF is an integer and CUTOFF >= 1,
then CUTOFF is considered as the maximum number of clusters to
keep in the hierarchical tree generated by LINKAGE.

The output T is a vector of size M that contains cluster number
for each observation.

T = CLUSTERDATA(X,CUTOFF) is the same as

```
Y = pdist(X,'euclid');
Z = linkage(Y, 'single');
T = cluster(Z, CUTOFF);
```

Follow this sequence to use non-default parameters for PDIST and
LINKAGE.

See also PDIST, LINKAGE, INCONSISTENT, CLUSTER.

ПРИЛОЖЕНИЕ В

Например, пусть загружены переменные A1, A2, A1v, A2v. Тогда выполнив команду A=[A1 A2 A1v A2v] получим матрицу A

```
128   123   103   128
128   122   103   131
129   122   103   133
130   123   103   135
130   123   103   136
129   122   105   137
130   122   107   138
129   121   108   138
129   122   109   137
130   121   111   136
130   121   111   135
129   121   112   134
..... и т.д.,
```

где столбец- номер QRS- комплекса, строка- значение отсчета в комплексе (первая строка- первый отсчет, вторая- второй и т.д.).

Эта матрица не готова для анализа с помощью команды **clusterdata.** Ее необходимо развернуть с помощью команды rot90. Например, B=rot90(A) даст следующий результат

Columns 1 through 18

128	131	133	135	136	137	138	138	137	136	135	134	132	130	128	127	125	125
103	103	103	103	103	105	107	108	109	111	111	112	112	112	111	112	112	112
123	122	122	123	123	122	122	121	122	121	121	121	123	125	127	128	131	132
128	128	129	130	130	129	130	129	129	130	130	129	131	134	136	138	140	142

Columns 19 through 36

124	123	123	123	123	123	123	123	123	123	120	115	102	87	77	79	86	91
111	111	111	110	110	111	110	110	109	107	105	98	71	31	3	3	3	25
131	129	128	126	124	122	122	122	122	122	120	126	158	213	230	174	109	92
141	141	140	135	130	128	127	128	127	126	124	127	152	209	239	190	123	105

Columns 37 through 54

96	102	107	111	112	113	113	113	113	114	115	117	119	121	122	124	125	127
58	90	108	117	121	125	127	130	132	133	134	137	139	140	140	139	139	140
110	117	120	119	120	120	121	120	121	121	122	122	123	124	124	125	126	127
120	123	123	123	122	123	122	123	123	123	124	125	126	126	127	128	128	128

Columns 55 through 64

129	132	135	137	140	142	144	146	147	146
143	146	148	152	156	161	165	169	171	174
127	128	131	133	135	136	138	141	144	147
130	132	134	135	138	140	143	145	148	149

Учебное издание

ИССЛЕДОВАНИЕ МЕТОДОВ И АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ БИОМЕДИЦИНСКИХ ДАННЫХ

Методические указания

Составитель: Конюхов Вадим Николаевич

Самарский государственный аэрокосмический университет
имени академика С.П. Королёва.
443086 Самара, Московское шоссе, 34