Лекция 12. Основы кластерного анализа в рамках статистического и детерминистского подходов

Задача кластеризации состоит в разбиении множества объектов, представленных своими образами, на кластеры, в каждом из которых объединяются в известном смысле «близкие» образы, в то время как образы, помещаемые в различные кластеры, имеют существенные «отличия».

Кластер (англ. cluster — гроздь, скопление) – группа однотипных объектов, имеющих определенную степень близость в пространстве используемых для их описания признаков.

Кластеризация по определению осуществляется в режиме самообучения (без учителя). В лучшем случае, изначально может быть задано количество кластеров, на которые следует разделить образы. Но часто и подобная информация отсутствует.

Существенно, что каждая группа должна содержать образы с такими общими признаками, которые позволяет рассматривать их как принадлежащие некоторому порождающему классу образов. Иными словами кластеры должны отражать сущности, обладающие общими категориальными свойствами. Поэтому эту задачу еще называют задачей классификацией без обучения.

Помимо этих терминов, в литературе встречается термин **таксономия**, который обозначает задачу разбиения и систематизации объектов по таксонам. Таксо́н (от др.-греч. «порядок, устройство, организация») — группа, состоящая из объектов, объединяемых на основании общих свойств и признаков.

Классифицирующие системы, использующие понятие «таксона», обычно носят *иерархический* характер, т.е. разбиение объектов на группы представляется в порядке от более мелких к более крупным (или наоборот). Суть задачи таксономии хорошо просматривается при систематизации объектов в биологии.

Задача обучения без учителя в интересах кластерного анализа данных, особенно в ситуации отсутствия априорной информации о числе кластеров, является, безусловно, сложнейшей в ряду всех рассмотренных. Она может решаться как в рамках статистического, так и в рамках детерминистского подходов.

1. Статистический подход к задаче классификации без обучения (EM – алгоритм)

В рамках статистического подхода рассматривается задача обучения без учителя, напрямую связанная с задачей кластеризации. Рассмотрим первоначально следующую постановку этой задачи.

1.Пусть имеется не помеченная, смешанная обучающая выборка образов $X^N = \{x^{(1)},...,x^{(N)}\}$, $x^{(*)} \in \mathbf{R}^n$, принадлежащих M различным классам, которые далее будем называть порождающими.

2. Число порождающих классов M считается известным, при этом априорные вероятности появления наблюдений каждого класса $\omega_i : \pi_i = p(\omega_i) > 0,\ i = \overline{1,M}$ неизвестны.

3. Считается известным или вводится обоснованнее предположение относительно вида плотностей распределения классов $\omega_i : p(x/\theta_i, \omega_i) = p(x/\theta_i),\ i = \overline{1,M}$, где $\theta_i \in \mathbf{R}^{h_i},\ i = \overline{1,M}$ векторы неизвестных параметров плотностей распределений классов.

Требуется оценить общий набор неизвестных параметров как составной вектор вида $\theta = (\theta_1^T,...,\theta_M^T,\pi_1...,\pi_M)^T$, т.е. произвести оценку параметров смеси распределений вида

$$p(x/\theta) = \sum_{i=1}^{M} p(x/\theta_i)\pi_i,\ i = \overline{1,M}\ . \qquad (1)$$

Далее по результатам оценки векторного параметра смеси необходимо произвести индексацию образов $X^N = \{x^{(1)},...,x^{(N)}\}$, привязав каждый из них к одной из компонент смеси, т.е. провести классификацию без учителя.

Задача может быть решена, если плотность распределения $p(x/\theta)$ в (1) существует и идентифицируема, т.е. если из $\theta \neq \theta'$ следует, что $p(x/\theta) \neq p(x/\theta')$. На практике большинство смесей идентифицируемы, в том числе и смеси гауссовских распределений, с которыми, в основном, и работают при проведении оценок параметров в задаче классификации без обучения.

Итак, пусть смесь идентифицируема. Представим функцию правдоподобия выборки в виде

$$p(X^N/\theta) = \prod_{k=1}^{N} p(x^{(k)}/\theta) = \prod_{k=1}^{N}\sum_{i=1}^{M} p(x^{(k)}/\theta_i)\pi_i\ .$$

Как и ранее, при решении задач оценивания параметров, задача может быть решена на основе двух базовых подходов, а именно: подхода, основанного на методе максимального правдоподобия, и байесовского подхода, базирующегося на методе максимума апостериорной вероятности.

При использовании **метода максимального правдоподобия** оцениваемые параметры рассматриваются как фиксированные и неизвестные.

При использовании **метода максимума апостериорной** вероятности оцениваемые параметры рассматриваются как случайные величины, для которых должно быть задано априорное распределение.

Мы остановимся на первом варианте решения задачи, учитывая, что получаемые результаты при применении указанных подходов в большинстве практически значимых случаев достаточно близки.

При использовании метода максимального правдоподобия оценка $\hat{\theta}$ формально может быть найдена как точка в многомерном пространстве, которая максимизирует логарифм функции правдоподобия

$$\tilde{\theta} = \arg\max_{\theta} \ln P(X^N / \theta), \quad \sum_{i=1}^{M} \pi_i = 1,$$

где сумма по неизвестным значениям априорных вероятностей классов играет роль ограничения к оптимизационной задаче. Решение подобной задачи в лоб в силу ее громоздкости практически невозможно.

Поэтому наибольшее распространение получил подход, основанный на реализации итерационных процедур поиска оптимального решения, которые называются алгоритмами максимизации математического ожидания или EM - алгоритмами (expectation - maximization). Особенностью этих алгоритмов является использование вспомогательной совокупности скрытых переменных Q. Скрытые переменные рассматриваются как наблюдаемые параметры и вводятся таким образом, чтобы их можно было вычислить при условии задания вектора $\theta = (\theta_1^T, ..., \theta_M^T, \pi_1 ..., \pi_M)^T$.

Это позволяет свести сложную оптимизационную задачу к последовательности итераций по пересчету скрытых переменных на основе текущего приближения вектора оцениваемых параметров (Е-шаг) и максимизации правдоподобия с целью найти следующее приближение (М-шаг).

Дополняющий вектор скрытых переменных вводится таким образом, чтобы на его основе можно было бы упростить максимизацию функции правдоподобия. Поскольку неизвестно, какой именно компонентой смеси распределений порожден каждый образ, для обозначения этого факта введем индексы $u_{ik}, i=\overline{1,M}, k=\overline{1,N}$, значения которых для нас скрыты. Здесь $u_{ik}=1$, если $x^{(k)} \in \omega_i$ и $u_{ik}=0$, если $x^{(k)} \notin \omega_i$. Именно эти индексы являются отражением скрытых параметров, оценку которых можно получить, зная компоненты смесей.

Удобно, в качестве скрытых переменных рассматривать апостериорные математические ожидания этих индексов. Они определяются вероятностями $p(\theta_i / x^{(k)})$ того, что образ $x^{(k)}$порожден классом ω_i (получен из компонента смеси с параметрами θ_i). Тогда соответствующее значение скрытой переменной при заданном значении θ рассчитывается как

$$Q=\{q_{ik}\}\ q_{ik}=\frac{\pi_i p(x^{(k)}/\theta_i)}{\sum_{i=1}^{M}\pi_i p(x^{(k)}/\theta_i)}, \quad \sum_{i=1}^{M} q_{ik}=1,\ \overline{k=1,N}, Q=\{q_{ik}\}. \tag{2}$$

Существо действий выполняемых на Е-шаге и М-шаге алгоритма наглядно можно пояснить на примере оценки смеси гауссовских распределений с неизвестными математическими ожиданиями и матрицами ковариациями $\omega_i : p(x/\theta_i, \omega_i) = N(x, m_i, C_i), i=\overline{1,M}$.

EM - алгоритм (стандартная реализация) по шагам:

1. Вводится начальное значение (начальное приближение) $\tilde{\theta} = (\tilde{\theta}_1^T, ..., \tilde{\theta}_M^T, \tilde{\pi}_1 ..., \tilde{\pi}_M)^T$, $\tilde{\theta}_i = \{\tilde{m}_i, \tilde{C}_i\}$, $i = \overline{1, M}$ и начальные значения $q_{ik} = 0$, $i = \overline{1, M}$, $\overline{k = 1, N}$. Для генерации начальной точки могут использоваться эвристические процедуры, основанные на любых разумных предположениях.

Далее в цикле, охватывающем последующие пункты, осуществляются следующие действия.

2. Выполняется Е-шаг, в ходе которого рассчитываются условные математические ожидания скрытых переменных (вероятностей их единичных значений)

$$q'_{ik} = \frac{\tilde{\pi}_i N(x^{(k)}, \tilde{m}_i, \tilde{C}_i)}{\sum_{i=1}^{M} \tilde{\pi}_i N(x^{(k)}, \tilde{m}_i, \tilde{C}_i)}, \quad i = \overline{1, M}, \ \overline{k = 1, N}.$$

3. Выполняется М-шаг, в ходе которого проводится перерасчет предыдущего приближения $\tilde{\theta} = (\tilde{\theta}_1^T, ..., \tilde{\theta}_M^T, \tilde{\pi}_1 ..., \tilde{\pi}_M)^T$ так, чтобы максимизировать правдоподобие наблюдений $\tilde{\theta} = \arg\max \ln P(X^N / \theta)$ при ограничении $\sum_{i=1}^{M} \pi_i = 1$. Перерасчет проводится с учетом вычисленных значений q'_{ik} следующим образом:

$$\tilde{\pi}'_i = \frac{1}{N} \sum_{k=1}^{N} q'_{ik} = \frac{N_i}{N}, \quad \tilde{m}'_i = \frac{1}{N_i} \sum_{k=1}^{N} q'_{ik} x^{(k)}, \ \tilde{C}'_i = \frac{1}{N_i} \sum_{k=1}^{N} q'_{ik} (x^{(k)} - \tilde{m}'_i)(x^{(k)} - \tilde{m}')^T.$$

4. Если разница значений скрытых переменных на соседних шагах $\delta_{\max} = \max |q_{ik} - q'_{ik}| > \delta_0$, то осуществляется присвоение $q_{ik} = q'_{ik}$, $i = \overline{1, M}$, $\overline{k = 1, N}$ $i = \overline{1, M}$, $\overline{k = 1, N}$ и переход на следующий шаг цикла в п. 2. Иначе – останов. В п.4. δ_0 – порог для фиксации останова итеративного процесса пересчета параметров при условии незначительности изменения скрытых переменных.

Можно показать, что вычисления, выполняемые на М-шаге, эквивалентны определению условия максимума логарифмов функций правдоподобия, рассчитываемых с весами q_{ik}, $i=\overline{1,M}$, $\overline{k=1,N}$

$$\tilde{\theta}_i = \arg\max_{\theta_i} \sum_{k=1}^{N} q_{ik} \ln p(x^{(k)} / \theta_i).$$

Также очевидно, что если апостериорные вероятности принимают значения 0 и 1, определяя однозначно принадлежность каждого образа тому или иному классу, то задача сводится к ранее рассмотренной задаче оценки параметров плотностей распределений классов по индексированным выборкам.

Качество решения и скорость сходимости ЕМ-алгоритма существенно зависят от начального приближения.

Стандартный прием, заключается в том, чтобы несколько раз повторить процесс оценивания, задавая каждый раз начальные значения параметров смеси случайным образом. В итоге выбирается тот результат, который будет наилучшим по заданному критерию, например, по максимуму правдоподобия наблюдений.

Другой прием состоит в том, чтобы взять в качестве центров компонент M объектов, максимально удалённых друг от друга, т.е. использовать максиминный алгоритм назначения центров.

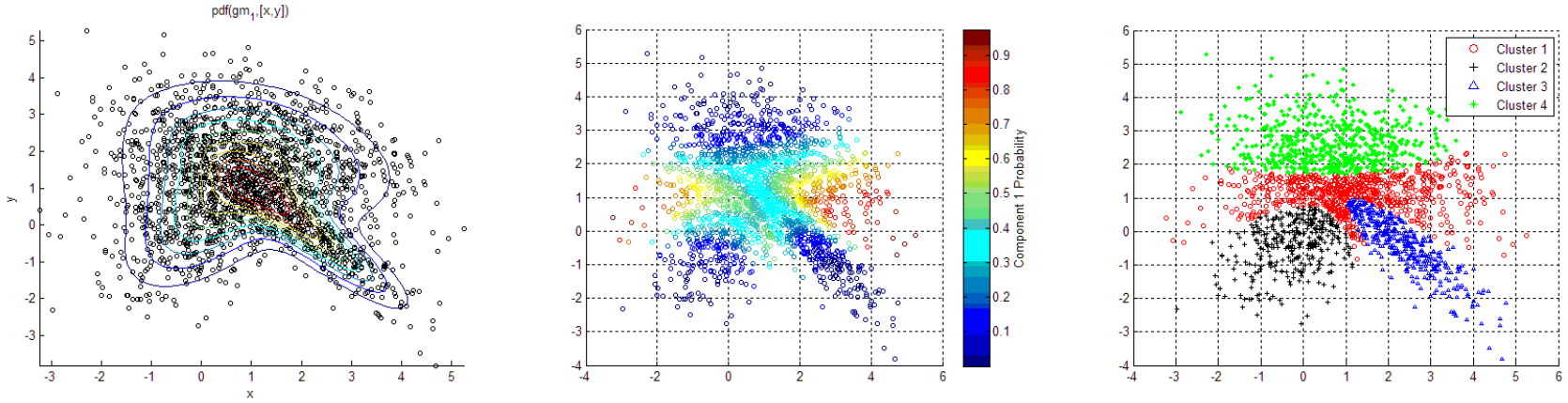
Он состоит в следующем. Первые два центра выделяются на основе нахождения максимума среди всех вычисляемых попарно расстояний между образами исходной выборки. Каждый следующий центр выбирается так, чтобы минимальное расстояние от назначаемого в качестве этого центра образа от него до ранее найденных центров было бы максимальным среди всех образов (образ является наиболее удаленным от ранее найденных центров).

Фактически в результате выполнения EM- алгоритма мы получаем весовые коэффициенты, характеризующие степень принадлежности образов тому или иному порождающему классу. Поэтому, в итоге, можно выполнить отнесение каждого образа обучающей выборки к группе, обозначаемой как $\tilde{\omega}_j$, в соответствии с максимальным значением

$$j = \arg\max_i q_{ik} \to x^{(k)} \in \tilde{\omega}_j,\ k = \overline{1,N}\ .$$

В ходе моделирования EM - алгоритма исследовался вопрос зависимости качества его работы от величины dm, характеризующей степень пересекаемости компонентов смесей, объема обучающих данных, а также от различных значениях объема выборки N. При проведении моделирования рассматривались два примера, отличающиеся вариантами задания исходных данных, определяющих конфигурацию кластеров: первый пример $n = 2$; $M = 4$, центры кластеров расположены в вершинах квадрата со сторонами dm; второй пример $n = 10$; $M = 3$, центры каждого кластера генерируются случайным образом по равномерному закону dm*rand(n,1) с фиксированным сдвигом dm=2.

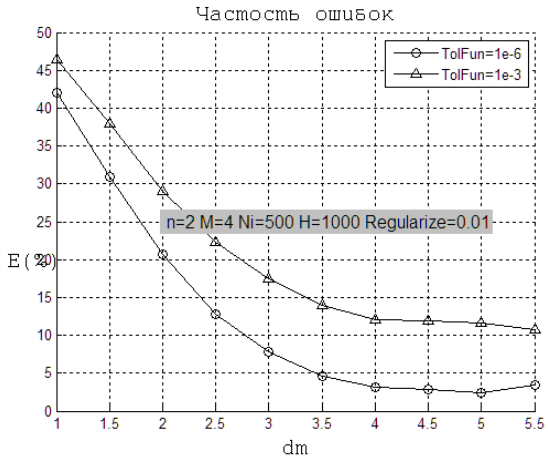
На рис.1 представлены результаты визуализации различных этапов анализа данных для первого примера. На рис.1a приведена локализация исходной смешанной выборки данных с наложенными на нее линиями постоянного уровня получаемой модели смеси, на рис.1b приведено отображение уровней вероятностей принадлежности данных компонентам смеси данных $Q = \{q_{ik}\}$, а на рис.1,в – результаты классификации, полученные при разметке данных.



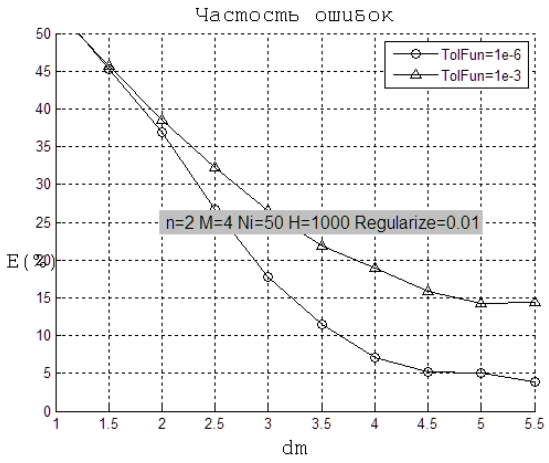
а) б) в)

Рис.1. Визуализация различных этапов анализа данных для 4 кластеров

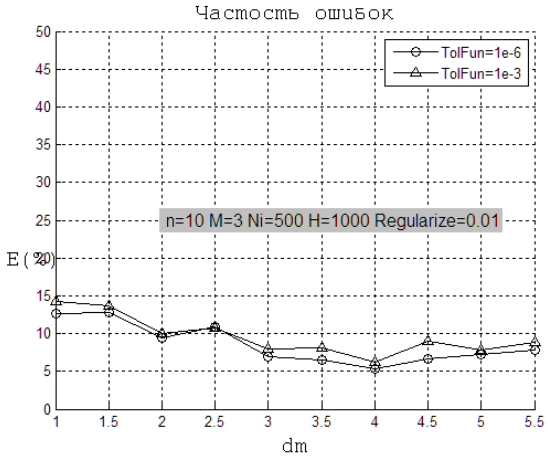
На рис.2,а,б,в,г приведены зависимости частости ошибок от величины dm для двух вариантов построения EM – алгоритма, отличающихся различной величиной порога, определяющего завершение итеративного процесса уточнения параметров смеси. Зависимости на рис.2,а,г представлены для первого примера задания данных при Ni=500 (а) и Ni=50 (б), определяющих конфигурацию кластеров. Зависимости на рис.2,в,г даны для второго примера задания данных, определяющих конфигурацию кластеров при тех же значениях количества образов, представляющих каждый кластер. Параметр H определяет количество проводимых экспериментов при тестировании и усреднении результатов кластеризации.



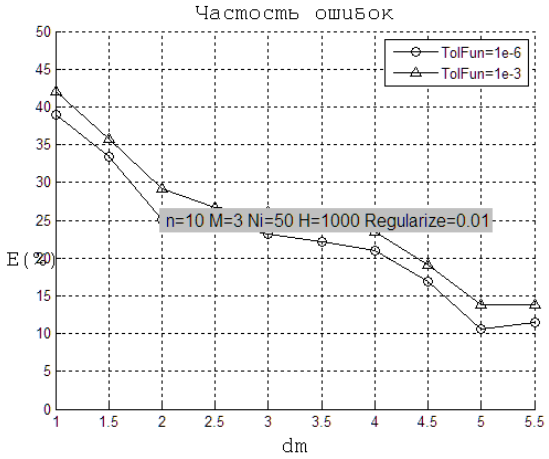
а)



б)



в)



г)

Рис.2. Результаты тестирования алгоритма для различных вариантов исходных данных

На приведенных графиках видно вполне естественное убывание уровня ошибок при увеличении параметра dm, которое в большей степени проявляется для первого примера. Следует также отметить, что при запуске программы при малых объемах выборки необходимо включать режим регуляризации, так как в этом случае велика вероятность получения в процессе выполнения итераций плохо обусловленных матриц.

Таким образом, мы видим, что в рамках статистического подхода задача кластеризации сводится к задаче оценивания параметров смеси распределений в гауссовском представлении. Такая кластеризация опирается на полученное решение, используя рассчитанные значения вероятностей принадлежности каждого образа выделенным компонентам смеси.

Исследованный здесь алгоритм не учитывает одну важную и часто встречающуюся на практике ситуацию: отсутствие информации о количестве порождающих классов. Для решения задачи в такой постановке существуют различные подходы, в том числе представленные в модификациях EM-алгоритма. Базовый подход основан на использовании специальных алгоритмов получения оценок числа классов при выполнении последовательности нескольких кластеризаций с разным значением M. Далее он будет рассмотрен и исследован в ходе сравнительного анализа различных алгоритмов, включая и EM- алгоритм.

2. Детерминистский подход к задаче классификации без обучения. Критерии оптимальности разбиения данных на кластеры

В рамках детерминистского подхода задача кластеризации формулируется следующим образом.

1. Имеется не помеченная, смешанная обучающая выборка образов $X^N = \{x^{(1)},...,x^{(N)}\}$, $x^{(*)} \in \mathbf{R}^n$, принадлежащих M различным порождающим классам.

2. Число порождающих классов M может быть, как известно, так и не известно.

3. Заданы функции расстояния (меры близости) $d(x^{(i)}, x^{(j)})$, дающие возможность судить о степени сходства образов в признаковом пространстве.

Требуется в соответствии с заданным критерием оптимальности выполнить разбиение $Q_M = \left\{X^{N_1},...,X^{N_M}\right\}$ смешанной выборки образов на группы $X^{N_i} = \{x^{(1,i)},...,x^{(N_i,i)}\}, i = \overline{1,M}$, обладающие следующими свойствами:

$$X^{N_i} \neq \varnothing, i = \overline{1,M}, \quad \bigcup_{i=1}^{M} X^{N_i} = X^N, \quad \bigcap_{i=1}^{M} X^{N_i} = \varnothing.$$

Следует сразу отметить, что такая постановка предусматривает реализацию исключительно однозначной (четкой) процедурой разбиения выборки на группы. Она является стандартной, но не единственной, так как в подобных задачах иногда лучше отказаться от однозначного ответа при привязке образа к кластеру (нечеткая кластеризация).

Тем не менее, далее мы будем придерживаться именно этой постановки задачи.

Решением задачи, как уже сказано, является такое разбиение, которое удовлетворяет некоторому критерию оптимальности. Этот критерий должен отражать уровень желательности различных вариантов разбиений и выполненной кластеризации в целом.

При не известном значении M единый строго обоснованный критерий оптимальности отсутствует и задача является некорректно поставленной. Это, однако, не означает, что нельзя использовать эвристические критерии, пригодные для получения решений в конкретных ситуациях. При обосновании подобных критериев могут использоваться различные показатели, вычисляемые для каждого возможного разбиения Q_M.

В этом плане, одним из основных показателей, используемых для определения качества кластеризации, является среднее суммы квадратов расстояния образов каждого кластера до его центра – внутриклассовый разброс (ВР), рассчитываемый как

$$E_W(X^N, Q_M) = \sum_{i=1}^{M} \sum_{k=1}^{N_i} d(x^{(k,i)}, m_i)^2, \quad m_i = \frac{1}{N_i} \sum_{k=1}^{N_i} x^{(k,i)}, i = \overline{1, M}.$$

При использовании в качестве функции расстояния Евклидовой метрики данная величина выражается через матрицы рассеяния классов

$$E_W(X^N, Q_M) = trS_W = tr\left(\sum_{i=1}^{M} S_i\right), \quad S_i = \sum_{k=1}^{N_i} (x^{(i,k)} - m_i)(x^{(i,k)} - m_i)^T,$$

где $S_i, i = \overline{1, M}$ – матрицы рассеяния образов каждого кластера.

Если число порождающих классов известно, оптимальным разбиением по данному показателю часто считается то, которое минимизирует

$$E_W(X^N, Q_M) \to \min .$$

Такой подход называется группированием с минимальной дисперсией. Данный критерий хорошо работает в случаях, когда порождающие классы образуют компактные, хорошо локализованные друг относительно друга кластеры.

При неизвестном числе классов величина $E_W(X^N, Q_M)$ будет монотонно уменьшаться с увеличением числа групп разбиения, доходя до нижней границы, равной нулю при разбиении на группы, содержащие по одному образу. Это означает, что напрямую этот показатель в данной ситуации не может быть использован.

Еще один важный показатель, фигурирующий в подобных задачах, основан на вычислении средней суммы квадратов расстояний между центрами кластеров относительно общего центра. Он характеризует межклассовый разброс (МР), рассчитываемый как

$$E_B(X^N, Q_M) = \sum_{i=1}^{M} N_i d(m_i, m)^2 , \quad m = \frac{1}{N} \sum_{k=1}^{N} x^{(k)} = \sum_{i=1}^{M} N_i m_i .$$

При использовании в качестве функции расстояния Евклидовой метрики МР выражается через матрицу рассеяния между классами

$$E_B(X^N, Q_M) = trS_B , \quad S_B = \sum_{i=1}^{M} N_i (m_i - m)(m_i - m)^T .$$

Принципиально то, что общая матрица рассеяния данных исходной выборки, независимо от разбиения, представляется в виде

$$S_T = \sum_{k=1}^{N} (x^{(k)} - m)(x^{(k)} - m)^T = \sum_{i=1}^{M} \sum_{k=1}^{N_i} (x^{(k,i)} - m)(x^{(k,i)} - m)^T =$$

$$= \sum_{i=1}^{M} \sum_{k=1}^{N_i} (x^{(k,i)} - m_i + m_i - m)(x^{(k,i)} - m_i + m_i - m)^T = \sum_{i=1}^{M} \left(S_i + N_i (m_i - m)(m_i - m)^T \right) =$$

$$= S_W + S_B .$$

Это означает, что применение любого разбиения на кластеры не изменяет величины следа общей матрицы рассеяния $J_T = trS_T = const$, которая в данном случае является инвариантом относительно разбиения Q_M.

Существование подобного инварианта позволяет надеяться на возможность формирования такого критерия качества кластеризации, который буде гарантировать определение разумного разбиения исходной выборки. Действительно, при увеличении количества групп величина $E_W(X^N, Q_M) = trS_W$ уменьшается, но при этом обязательно будет увеличиваться величина $E_B(X^N, Q_M) = trS_B$. Это потенциально позволит регуляризировать решение и избежать появления заведомо неприемлемого.

Существует много подходов и конкретных алгоритмов кластеризации, которые подробно рассмотрены в ранее упомянутых источниках. К ним относятся, например, подходы, основанные на выполнении итеративной оптимизации процесса разбиения данных на группы; подходы, основанные на иерархическом представлении процесса формирования групп; подходы, опирающиеся на графовые описания задачи, и многие другие. Далее мы ограничимся рассмотрением двух базовых алгоритмов в условиях известного и не неизвестного числа классов.

3. Кластеризация при известном числе классов. Алгоритм K- внутригрупповых средних и алгоритм иерархической кластеризации

Рассмотрим ситуацию, когда число классов заранее известно. В этом случае задача становится корректно поставленной задачей дискретной оптимизации. При ее решении требуется перебрать все возможные комбинации разбиения и использовать критерий, отражающий качество получаемого результата, например, минимума внутриклассового разброса. Очевидно, что прямой перебор всех комбинаций слишком затратен и нужен другой, более разумный способ нахождений решения, которое будет близок к оптимальному. Рассмотренные ниже алгоритмы и представляют два возможных альтернативных варианта решения этой задачи.

Алгоритм K-внутригрупповых средних (K – means). Далее имеется в виду, что $K = M$ – заданное число кластеров. В основе алгоритма лежит процедура объединения данных в группы с минимальной дисперсией, т.е. минимизация показателя $E_W(X^N, Q_M)$. Алгоритм реализует итеративную оптимизацию критерия, которая отражает процесс направленного поиска решения.

Следует отметить, что при таком поиске глобально оптимальное решение может быть не найдено, а вместо него может быть получено локально оптимальное решение, т.е. локальный минимум функции критерия.

Идея алгоритма состоит в нахождении некоторого приемлемого начального разбиения данных на группы и выполнении нескольких итераций, на каждой из которых производится передвижение образов из одной группы в другую с одновременным пересчетом функции критерия и проведением контроля за изменениями ее значения. При этом на каждом шаге сохраняются такие перемещения, которые приводят к уменьшению значения $E_W(X^N, Q_M)$.

Алгоритм K – means (стандартная реализация) по шагам:

1. Выбираются начальные значения M центров кластеров $m_i, i=\overline{1,M}$, в качестве которых назначаются либо первые образы из исходной выборки X^N (т.е. фактически случайно), либо M наиболее удаленных друг от друга образов выборки (максиминный алгоритм).

Далее в цикле, охватывающем последующие пункты, осуществляются следующие действия.

2. Проводится разбиение $Q_M = \{X^{N_1},...,X^{N_M}\}$ исходной выборки на кластеры, в ходе которого принадлежность каждого образа определяется на основе поиска минимума среди расстояний до установленных ранее центров m_i, $i=\overline{1,M}$. Рассчитывается начальное значение $E_W(X^N, Q_M)$.

3. Проводится перерасчет центров с учетом результатов выполненной в п.2 кластеризации

$$m_i' = \frac{1}{N_i}\sum_{k=1}^{N_i} x^{(k,i)}, i=\overline{1,M}$$

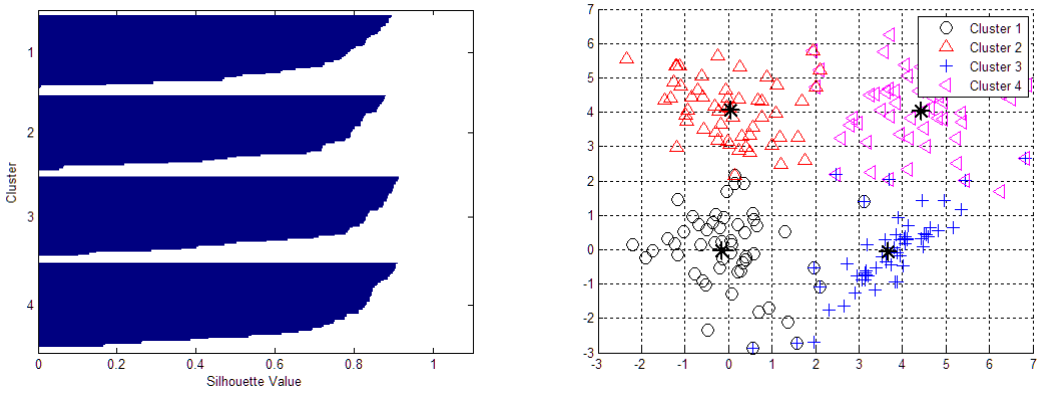
и соответствующий перерасчет функции $E'_W(X^N, Q_M)$.

4. Если разница значений центров на соседних шагах $\delta_{\max} = \max|m_i - m_i'| > \delta_0$, то осуществляется присвоение $m_i = m'_i$, $i=\overline{1,M}$ и переход к выполнению следующего шага в п. 2. Иначе – останов.

В п.4. δ_0 – порог для фиксации останова итеративного процесса пересчета при условии незначительности изменения центров кластеров на соседних шагах. При реализации подобной процедуры значения функции $E'_W(X^N, Q_M) \le E_W(X^N, Q_M)$ не увеличиваются на соседних шагах и алгоритм сходится за конечное число шагов. Это означает, что в процессе выполнения алгоритма выполняется **направленный поиск решения**.

Один из приемов, который может повысить устойчивость алгоритма при случайном начальном назначении центров, состоит в выполнении неоднократного запуска со случайными начальными значениями и выборе того решения, которое покажет минимум функции критерия. Тем самым можно повысить устойчивость алгоритма по отношению к начальному назначению центров кластеров.

Для визуализации результатов кластеризации здесь используется «силуэт» результатов, который дает представление о распределении величины, равной отношению разности среднего расстояние до членов ближайшего кластера и среднее расстояние до членов своего кластера, деленного на максимум из этих двух средних. Значение силуэта изменяется от –1 до +1. Чем дальше будут образы, находящиеся в соседнем кластере, тем ближе к 1 значение силуэта.



а) б)

Рис.3. Пример отображения силуэта исходных кластеров и полученных решений (dm=3,Ni=50)

Алгоритмы иерархической кластеризации. Общая идея этих алгоритмов состоит в последовательном объединении или разъединении групп образов, для которых установлены определенные правила слияния групп или, соответственно, правила расщепления. При достижении заданного количества групп (или необходимого уровня другого критерия) процесс останавливают. При объединении групп процесс называется **агломеративной** кластеризацией, а при разъединении – **дивизимной** кластеризацией.

Рассмотрим последовательность разделений N образов исходной выборки X^N на M групп следующим образом:

- первое разделение Q_N^1 состоит из N групп, содержащих по одному образу в каждой;
- второе разделение Q_{N-1}^2 состоит из $N-1$ групп, содержащих по одному образу в $N-2$ группах и два образа в одной группе;

  …
- k - ое разделение Q_{N-k+1}^k состоит из $N-k+1$ групп, содержащих один или несколько образов в каждой;

  ...
- N - ое разделение Q_1^N состоит из одной группы, содержащей N образов.

Говорят, что процесс обработки данных находится на i - ом уровне в этой последовательности, если выделенное количество групп (кластеров) $M = N-k+1$. Порядок расположения уровней будем определять в соответствие с их номерами (меньшие номера внизу).

Очевидно, что подобная последовательность разделений может быть получена как при движении снизу вверх путем объединения групп (агломеративное), так и при движении сверху вниз путем расщепления групп (дивизимное).

Главная идея организации такого процесса состоит в том, два любых образа на некотором уровне обязательно находятся в одной группе.

Если при агломеративной кластеризации два образа попадают на каком-либо уровне в одну группу, то они в ней и останутся на более высоких уровнях.

Если при дивизимной кластеризации два образа находятся на каком-либо уровне в одну группу, то они обязательно окажутся в разных группах на более низких уровнях.

Такой подход обеспечивает направленный перебор комбинаций при кластеризации образов с постепенным улучшением используемых критериев.

Особенностью иерархических алгоритмов является то, что при их реализации обычно строят полное дерево вложенных кластеров – дендрограмму.

Агломеративный алгоритм кластеризации (стандартная реализация) по шагам:

1. Устанавливается значение $\tilde{M} = N$ и формируется начальное разбиение на группы (начальная кластеризация) Q_N^1, $Q_N^1 = \{X^1,...,X^N\}$, $X^j = x^{(j)}$.

2. В цикле для выполненной ранее кластеризации Q_{N-k+1}^k и $\tilde{M} = N-k+1$ с использованием заданной функции расстояния групп находится ближайшая пара групп, имеющая $\min\limits_{i,j} D(X^{N_i}, X^{N_j})$.

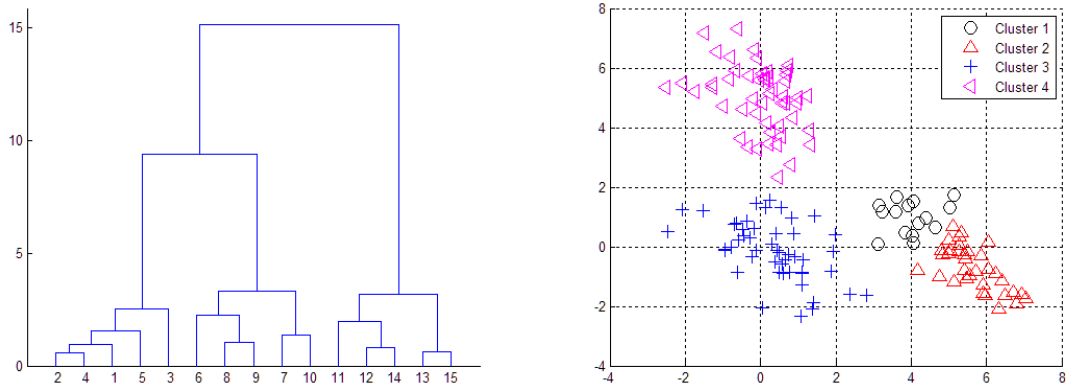
3. Проводится слияние выбранных групп $X^{N_i+N_j} = X^{N_i} \bigcup X^{N_j}$, изъятие X^{N_j} из Q_{N-k+1}^k и образование новой кластеризации Q_{N-k}^k с соответствующими новыми обозначениями групп и присвоением $\tilde{M} = N-k$.

4. Если $\tilde{M} = 1$ (можно $\tilde{M} = M$, если число классов заранее задано) – останов.

Принципиальное значение для подобных алгоритмов имеет выбор функции расстояния для определения близости кластеров:

- расстояние между ближайшими соседями $D^{NN}(X^{N_i}, X^{N_j}) = \min\limits_{k,s} d(x^{(k,i)}, x^{(s,j)})$;
- расстояние между самыми удаленными соседями $D^{FN}(X^{N_i}, X^{N_j}) = \max\limits_{k,s} d(x^{(k,i)}, x^{(s,j)})$;
- среднее расстояние между образами кластеров $D^{BG}(X^{N_i}, X^{N_j}) = \dfrac{1}{N_i N_j} \sum\limits_{k=1}^{N_i} \sum\limits_{s=1}^{N_j} d(x^{(k,i)}, x^{(s,j)})$;
- расстояние между центрами кластеров $D^{BC}(X^{N_i}, X^{N_j}) = d^2(m_i, m_j)$, $m_i = \dfrac{1}{N_i} \sum\limits_{k=1}^{N_i} x^{(k,i)}$, $m_j = \dfrac{1}{N_j} \sum\limits_{s=1}^{N_j} x^{(s,i)}$;

- расстояние Уорда (минимум соответствует максиму прироста суммы квадратов отклонений до центров кластеров при их объединении $D^{WM}(X^{N_i}, X^{N_j}) = \frac{N_i N_j}{N_i + N_j} d^2(m_i, m_j)$.



а), б)

Рис.4. Отображение полной дендрограммы при M=3,Ni=5 (а), а также получаемых кластеров при завышенном на единицу значении 'maxclust' (M=3,dm=4,Ni=50)

4. Критерии оценки числа классов и сравнительный анализ алгоритмов кластеризации в условиях неизвестного числа классов

Рассмотри, наконец, самую «тяжелую» в задачах кластерного анализа ситуацию, когда число классов неизвестно. Основной подход – последовательное выполнение нескольких кластеризаций с перебором числа порождающих классов $\tilde{M}$. В итоге осуществляется выбор того варианта с оценкой числа классов, который является наилучшим по определенному критерию. При оценке числа классов по результатам выполнения нескольких вариантов кластеризации с переменным значением $\tilde{M}$ могут с различным успехом использоваться несколько критериев оптимальности кластеризации.

1. Критерий (индекс) Калинского - Харабаша. Основан на максимизации отношения показателей межклассового и внутриклассового разброса при переборе нескольких кластеризаций. Рассчитывается с использованием ранее веденных показателей BP и MP по следующей формуле:

$$I_{CH}(\tilde{M}) = \frac{E_B(X^N, Q_{\tilde{M}})(N - \tilde{M})}{E_W(X^N, Q_{\tilde{M}})(\tilde{M} - 1)}.$$

2. Критерий (индекс) Хржановского – Лаи. Основан на выявлении максимального скачка BP при последовательном изменении числа классов. Рассчитывается по следующей формуле:

$$I_{CL}(\tilde{M}) = \left| \frac{\Delta E_W(\tilde{M})}{\Delta E_W(\tilde{M}+1)} \right|, \quad \Delta E_W(\tilde{M}) = (M-1)^{1/n} E_W(X^N, Q_{\tilde{M}-1}) - M^{1/n} E_W(X^N, Q_{\tilde{M}}).$$

3. Критерий (индекс) Дэвиса - Болдуина. При его вычислении определяют среднюю схожесть между каждым кластером и наиболее близким ему кластером. На каждом шаге изменения $\tilde{M}$ проводятся вычисления степени схожести образов в каждом кластере и степени различия между кластерами в целом:

$$e_{wi} = \left[\frac{1}{N_i} \sum_{k=1}^{N_i} \left\| x^{(k,i)} - m_i \right\|^p \right]^{1/p}, h_{ij} = \left[\sum_{k=1}^{n} \left| m_{i,k} - m_{j,k} \right|^q \right]^{1/q}.$$

При вычислении этих показателей в стандартном варианте используется Евклидово расстояние. В итоге индекс вычисляется по формуле:

$$I_{DB}(\tilde{M}) = \frac{1}{\tilde{M}} \sum_{i=1}^{\tilde{M}} R_i, \quad R_i = \max_j R_{ij}, \quad R_{ij} = (e_{wi} + e_{wj}) / h_{ij},$$

где R_{ij} – мера схожести двух кластеров.

4. Индекс GAP. Основан на анализе статистики расхождений (gap – расхождение, зазор, скачок) между вычисленным значением ВР $E_W(X^N, Q_{\tilde{M}})$ и усредненными значениями этой же величины, полученными при кластеризации унифицированных выборок, сгенерированных методом Монте-Карло на основе некоторого стандартного распределения

$$I_{GAP}(\tilde{M}) = \frac{1}{B} \sum_{k=1}^{B} \log E_W^*(X^N, Q_{\tilde{M}}) - \log E_W(X^N, Q_{\tilde{M}}).$$

Особенностью данного показателя является возможность вычисления даже в случае одного кластера.

5. Индекс оценки силуэта. Основан на вычислении величины «силуэта» для каждого образа, который определяет, насколько этот образ схож с образами собственного кластера, и, как он отличается от образов других кластеров. Индивидуальный индекс силуэта для образа $x^{(k,i)} \in X^{N_i}$ рассчитывается как отношение вида

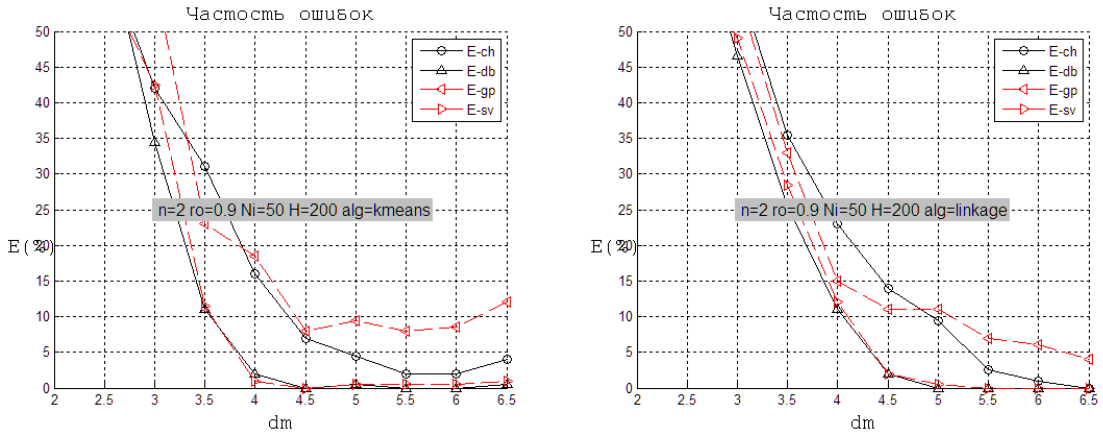
$$r_{i,k} = (b_{i,k} - a_{i,k}) / \max(b_{i,k}, a_{i,k}),$$

$$a_{i,k} = \frac{1}{N_i - 1} \sum_{s=1}^{N_i} d(x^{(k,i)}, x^{(k,s)}), \quad b_{i,k} = \min_{j \neq i} \left[\frac{1}{N_j} \sum_{t=1}^{N_j} d(x^{(k,i)}, x^{(t,j)}) \right],$$

где $a_{i,k}$ среднее расстояние от $x^{(k,i)}$ до образов своего класса; $b_{i,k}$ минимальное среднее расстояние от $x^{(k,i)}$ до образов других классов. В итоге вычисляется суммарный силуэт, максимальное значение которого при переборе нескольких кластеризаций дает оценку числа классов:

$$I_{SV} = \frac{1}{N} \sum_{i=1}^{\tilde{M}} \sum_{k=1}^{N_i} r_{i,k} = (b_{i,k} - a_{i,k}) / \max(b_{i,k}, a_{i,k}).$$

На рис.5а,б приведены зависимости частости ошибок при оценке числа классов от величины dm для двух типов алгоритмов кластеризации, используемых при переборе гипотез относительно числа классов. Зависимости на рис.5a,b представлены при $n = 2$ для величины Ni=50, определяющей минимальный объем образов, генерируемых для каждого порождающего класса. Из приведенных графиков видно, что лучшие в данном примере результаты показывают критерий Дэвиса – Болдуина и индекс силуэта. При этом более эффективным является использование указанных показателей совместно с алгоритмом кластеризации K- means (K-внутригрупповых средних). В этом случае пороговое значение величины dm, при котором уровень ошибки становится меньше 5%, будет меньше почти на единицу.



а) б)

Рис.5. Зависимости для частости ошибок при использовании различных критериев оценки числа классов

На рис.6. представлены результаты моделирования трех типов ранее рассмотренных алгоритмов кластеризации Cluster with Gaussian Mixtures (EM - алгоритм), k-Means Clustering, Hierarchical Clustering и оценивается качество их работы при не известном заранее числе порождающих классов. При этом сначала проводится оценка числа классов на основе одного из представленных критериев (в примере на основе критерия Дейвиса – Болдуина), а затем полученная оценка подставляется как входной параметр алгоритм кластеризации. При проведении моделирования рассматривался вариант задания исходных данных, определяющих конфигурацию кластеров для $n=2$. Случайным образом проводилась генерация от 2 до 7 порождающих классов с центрами, расположенными в точках, отличающихся координатами по первым двум признакам на величину dm.

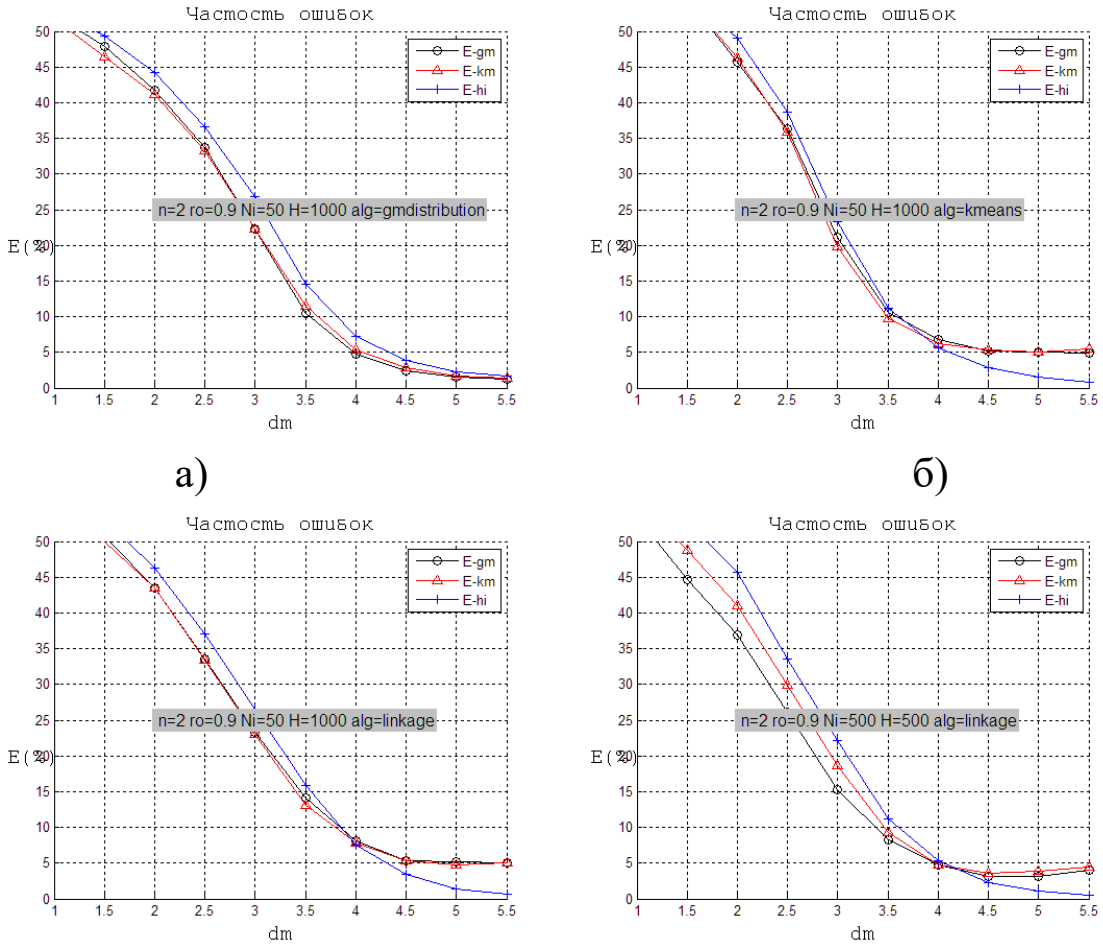


Рис.6. Типовые зависимости для частости ошибок при моделировании вариантов совместного применения алгоритмов оценки числа классов и окончательной кластеризации

Из трех используемых алгоритмов в примере, представленном на рис.6, на этапе окончательной кластеризации при малых и средних значениях параметра dm наилучшие результаты показывает EM-алгоритм. Ему незначительно уступает алгоритм K - средних. Алгоритм иерархической кластеризации ('linkage') проигрывает им при малых и средних значениях dm, но устойчиво показывает лучшие результаты при высокой степени разделимости классов (когда dm больше 4).

Таким образом, мы видим, что задача кластерного анализа при неизвестном числе классов имеет неоднозначное решение с точки зрения выбора используемых алгоритмов и рекомендаций по их настройкам.

Это определяет исключительную важность исследований, направленных на моделирование и сравнительный анализ альтернативных вариантов используемых методов и алгоритмов кластеризации применительно к конкретным ситуациям.

Подобный вывод можно сделать по отношению к любым из ранее рассмотренных методов и алгоритмов анализа данных.

Рассматривая разнообразные методы и алгоритмы анализа данных, мы часто обращали внимание на системные характеристики, напрямую влияющие на их качество и возможности практического применения.

К этим характеристикам относятся такие, как: точность алгоритма анализа данных (измеряемый уровень допущенных ошибок), вычислительная сложность и быстродействие, устойчивость по отношению к аномальным наблюдениям, способность к обобщению – сохранению уровня допускаемых ошибок для данных, не участвовавших в процессе обучения, сложность настройки параметров алгоритма, наличие программных библиотечных компонентов для его реализации.

Учитывая противоречивый характер требований к общей эффективности применяемого алгоритма, вытекающих из необходимости обеспечения указанных качеств и свойств, задача выбора подходящего варианта становится для исследователя и разработчика нетривиальной.

Решить ее можно только в ходе проведения масштабных и многофакторных модельных экспериментов, выполняемых для «пилотных» вариантов алгоритмов, с применением для этих целей современных инструментальных средств компьютерного моделирования