Некоторые алгоритмы классификации, основанные на евклидовой метрике. Оценка результатов процесса кластеризации

- Расстояние между эталонами
- Оценка дисперсии
- Информация о наиболее близкого и удаленного от эталона объекта

Приведем краткий обзор методов кластерного анализа, основанных на квадрате евклидова расстояния между S_i и S_j

$$^2=(S_i - S_j)^T(S_i - S_j). \qquad (2.22)$$

В работе [17] Соренсен описывает метод полных связей. Суть этого метода заключается в том, что два объекта, принадлежащие одному и тому же кластеру , имеют коэффициент сходства, который меньше некоторого порогового значения. Это означает, что расстояние между двумя объектами кластера не должно превышать некоторого порогового значения *r*, т. е. *r* - максимально допустимый диаметр подмножества - кластера.

Ворд [17] в качестве целевой функции применяет внутригрупповую сумму квадратов отклонений, которая есть не что иное, как сумма квадратов расстояний между каждым объектом и средним значением по кластеру, содержащему этот объект. Его метод также представляет собой последовательную процедуру. На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, определенной как внутригрупповая сумма квадратов отклонений. Такой метод направлен на объединение близко расположенных кластеров.

Сокал и Мигенер [43] описывают процедуру , которую назвали центроидным методом. Расстояние между кластерами K_i и K_j в этом случае определяется как евклидово расстояние между центрами (средними) этих кластеров, как

$$\rho(K_i, K_j)=\left(\bar{S}^i - \bar{S}^j\right)^T\left(\bar{S}^i - \bar{S}^j\right) \qquad (2.23)$$

где $\bar{S}^i, \bar{S}^j$ - обозначают векторы средних, соответственно, кластеров K_i, K_j.

Кластеризация осуществляется поэтапно: на каждом из *m-1* шагов объединяют два кластера K_i, K_j, имеющие минимальное значение расстояния (K_i, K_j).

В работе [43] Сокал предлагает другой метод, который называется двухгрупповым и опирается на связь между объектом S_i и кластером K_j. Эта связь выражается в виде среднего коэффициента сходства между объектом S_i и всеми объектами, входящими в кластер K_j. Для того чтобы средний коэффициент сходства выразить через евклидово расстояние, обозначим объекты, входящие в кластер K_j, соответственно через $S_1^j, S_2^j, \ldots, S_{\text{mj}}^j$, а через $\bar{S}^j$- центр кластера K_j. Тогда среднее расстояние (S_i, K_j) между объектами S_i K_j и всеми объектами из K_j будет следующим:

$$\rho(S_i, K_j)=\frac{1}{m_j}\sum_{v=1}^{m_j} ¿¿. \qquad (2.24)$$

Преобразовывая, получим двухслагаемое выражение

$$\rho(S_i, K_j)=\frac{1}{m_j}\sum_{v=1}^{m_j} ¿¿ \qquad (2.25)$$

¿

Первое слагаемое правой части уравнения называется внутригрупповой дисперсией объектов из K_j; второе слагаемое представляет собой квадрат расстояния между S_i и центром кластера K_j. Процедура последовательной кластеризации заключается в том, что объект S_i K_j, для которого (S_i, K_j) минимально, присоединяется к кластеру K_j.

Ланс и Уильямс [17] обобщают двухгрупповой метод и определяют среднее сходство между двумя кластерами K_i и K_j как среднее сходство между всеми парами объектов из K_i и

K_j. Процедура последовательная, два кластера с минимальным средним коэффициентом сходства объединяются.

В работах [24, 29] описывается метод, в котором объект, служащий начальной точкой, выбирается случайно. Все объекты, лежащие на расстоянии от начальной выбранной точки не больше *r*, принимаются как объекты первого кластера. Из оставшихся объектов выбирается второй центр-объект и процесс повторяется. В результате все группы будут развиты как кластеры.

Конечно, качество алгоритма зависит от последовательности выбора центральных объектов (эталонов) и от значения радиуса *r*. Выбор радиуса основывается на результатах предварительной выборки.

В методе Болла и Холла [51] первоначально формируется *l* центров кластеров случайным образом, к которым затем присоединяется каждый из оставшихся *m-l* объектов - по минимальному расстоянию к той или иной из них. Затем находятся центры кластеров и два кластера K_i, K_j объединяются, если (K_i, K_j) меньше некоторого порогового значения *r*. Наоборот, если внутригрупповая дисперсия кластера по некоторой переменной превосходит пороговое значение, то кластер разбивается.

Теперь рассмотрим наиболее популярные алгоритмы кластерного анализа, которые часто встречаются при исследовании проблем распознавания образов и кластерного анализа, а также при решении задач в различных отраслях науки и технологических процессов.

Принципиальная трудность оценки результатов алгоритмов кластеризации связана с тем, что мы не в состоянии зрительно представить геометрические особенности многомерного пространства. Хотя в предыдущих примерах число измерений было ограничено двумя с тем, чтобы облегчить изложение основ метода, читатель должен иметь в виду, что в большинстве задач распознавания образов размерность много выше. Поэтому для того, чтобы иметь возможность должным образом интерпретировать результаты процедуры отыскания кластеров, нам следует обратиться к схемам, которые обеспечивают по крайней мере некоторое представление о геометрических свойствах полученных кластеров. Ниже описывается несколько методов интерпретации результатов кластеризации.

При интерпретации очень полезно использовать расстояние между центрами кластеров. Лучше всего информацию подобного рода представлять с помощью таблиц типа табл. 2.1, составленной для модельного численного примера; из которой можно почерпнуть ряд важных сведений.

Таблица 2.1.

Пример таблицы расстояний для интерпретации результатов кластеризации

Центры кластеров	z_1	z_2	z_3	z_4	z_5
z_1	0	4,8	14,7	2,1	50,6
z_2	,0	0,0	21,1	6,1	48,3
z_3			0,0	15,0	36,7
z_4				0,0	49,3
z_5					0,0

Наиболее важным является то обстоятельство, что центр кластера z_5 существенно смещен относительно четырех других центров кластеров. Кроме того, расстояния между

центрами кластеров z_1 и z_2, как, впрочем, между z_1 и z_4 относительно одинаковы, если разделять только близко и далеко расположенные центры кластеров.

Таблица расстояний не является, естественно, достаточной основой для получения содержательных выводов. При интерпретации таблицы расстояний обычно используют в качестве вспомогательного средства количество образов классифицируемой выборки, вошедшее в каждый кластер. Так, например, из табл. 2.1 следует, что центр кластера z_5 далеко отстоит от центров остальных кластеров. Если известно, что в этот кластер входит много образов, его следует принять в качестве элемента истинного описания данных. Если же, с другой стороны, в кластер входит только один или два образа, можно после соответствующего анализа устранить этот центр кластера, заключив, что данные образы являются шумом. Может, естественно, оказаться, что образ, сильно отличающийся от всех других, представляет существенное событие, но установить это позволит лишь скрупулезный анализ представленных данных.

Информацию об образах, содержащихся в кластерах, можно также использовать при проведении объединения кластеров. Если центры двух кластеров расположены сравнительно близко друг от друга и в одном из соответствующих кластеров содержится намного больше образов, чем в другом, то часто удается слить эти кластеры в один.

Рассеяние характеристик кластера относительно средних значений можно использовать для получения представления об относительном расположении образов внутри кластера. Эту информацию легко оформить в виде таблицы дисперсий типа табл. 2.2, построенной для модельного примера (для простоты принято, что образы четырехмерные).

Таблица 2.2.

Пример таблицы дисперсий для интерпретации результатов кластеризации

Кластеры	Дисперсия			
	δ_1	δ_2	δ_3	δ_4
S_1	1,2	0,9	0,7	1,0
S_2	2,0	1,3	1,5	0,9
S_3	3,7	4,8	7,3	10,4
S_4	0,3	0,8	0,7	1,1
S_5	4,2	5,4	18,3	3,3

Как и раньше, S_i обозначает i кластер. Мы считаем, что каждая компонента дисперсии представляет отклонение по одной из координатных осей. На основании табл. 2.2 можно установить некоторые свойства классифицируемой выборки образов. Так, поскольку кластер S_1 характеризуется примерно одинаковыми дисперсиями по всем осям координат, можно предположить, что его форма близка к сферической. С другой стороны, кластер S_5 отличается значительной протяженностью вдоль третьей оси координат. Подобным же образом можно проанализировать и остальные кластеры. Эта информация в сочетании с таблицей

расстояний и списком образов, входящих в каждый из выделенных кластеров, может оказаться весьма ценным подспорьем при интерпретации результатов кластеризации.

Естественно, существует множество других количественных оценок кластерной структуры. Полезно, например, иметь сведения о ближайшей и наиболее удаленной от центра кластера точках для всех кластеров. Помимо информации, содержащейся в таблице расстояний, можно учитывать среднюю величину расстояния между центрами кластеров. Ковариационная матрица, построенная для множества образов каждого кластера, также представляет значительный интерес, хотя в задачах затруднений высокой размерности ее непросто интерпретировать, а вычисление может вызвать затруднения при реализации итеративного алгоритма.

При использовании оценок качества кластеризации, типа приведенных выше, информацию следует представлять в таком виде, чтобы соответствующая интерпретация не вызвала неопределенностей. Поскольку эта информация часто используется для коррекции выбора параметров в процессе выполнения итеративного алгоритма (например, алгоритма ИСОМАД), принято встраивать в соответствующие процедуры операции, связанные с вычислением и воспроизведением выбранного набора оценок качества кластеризации. Характер алгоритмов отыскания кластеров показывает, что наилучший способ их реализации - режим диалога, когда результаты каждого цикла итерации представляются пользователю в таком виде, чтобы он, выбирая нужные параметры, мог управлять процессом выполнения алгоритма.

Проблема выбора наилучшей кластеризации на основе выбранного критерия качества является самостоятельной задачей. Критерий кластеризации может либо воспроизводить некие эвристические соображения, либо основываться на минимизации (или максимизации) некоторого показателя качества.

Различные приложения методов кластерного анализа оставляют свои влияния на вид выбираемых показателей критерия качества.

Одним из популярных показателей является сумма квадратов ошибки

$F=\sum_{j=1}^{l}\sum_{S\in K_j}\left\|S-z_j\right\|^2$, где l - число кластеров; K_j - кластер; z_j - вектор выборочных средних для кластера K_j (или эталон).

Другие критерии качества приведены в §2.7 и §2.11 в данной книге.