§ 1.2. Многомерный кластерный анализ

**Задача кластерного анализа.** Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве , разбить множество объектов на ( – целое) кластеров (подмножеств) , , …, , так, чтобы каждый объект принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие одному и тому же кластеру, были сходными, в то время, как объекты, принадлежащие разным кластерам были разнородными (см. [9],[10], [11], [12], [15] и др.).

Решением задачи кластерного анализа являются разбиения, удовлетворяющие некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, который называют целевой функцией. Например, в качестве целевой функции может быть взята внутригрупповая сумма квадратов отклонения:

,

где - представляет собой измерения -го объекта.

Для решения  задачи кластерного анализа необходимо определить понятие сходства и разнородности.

Объекты -ый и -ый попадут в один кластер, когда расстояние между точками и будет достаточно мало и попадут в разные кластеры, когда это расстояние будет достаточно большим. Попадание в один или разные кластеры объектов определяется понятием расстояния между и из , где – -мерное евклидово пространство.

Неотрицательная функция называется функцией расстояния (метрикой), если:

а) для всех и из ;

б) тогда и только тогда, когда ;

в) ;

г) ;

где – любые три вектора из .

Наиболее часто употребляются следующие функции расстояний:

евклидово расстояние        ,

– норма                            ,

супремум - норма              , ,

– норма                             .

Евклидова метрика является наиболее популярной. Метрика наиболее легкая для вычислений. Супремум-норма легко считается и включает в себя процедуру упорядочения, а - норма обобщает три первые функции расстояний.

Пусть n измерений , ,..., представлены в виде матрицы данных размером :

.

Тогда расстояние между парами векторов могут быть представлены в виде симметричной матрицы расстояний:

.

Важным понятием является понятие сходства между объектами и . Неотрицательная вещественная функция  называется мерой сходства, если:

1) для ;

2) ;

3) ;

Пары значений мер сходства можно объединить в матрицу сходства:

.

Величину называют коэффициентом сходства.

Естественной  мерой сходства характеристик объектов во многих задачах является коэффициент корреляции между ними



где , , , соответственно средние и среднеквадратичные отклонения для характеристик и . Мерой различия между характеристиками может служить величина  . В некоторых задачах  знак коэффициента корреляции несуществен и зависит лишь от  выбора единицы измерения. В этом случае в качестве меры различия  между характеристиками используется .

**Методы кластерного анализа.** Методы кластеризации могут быть иерархическими и неиерархическими. Иерархическая кластеризация (hierarchical clustering) характеризуется построением иерархической, или древовидной, структуры.

Иерархические методы могут быть агломеративными (объединительными) и дивизивными. Агломеративная кластеризация (agglomerative clustering) начинается с каждого объекта в отдельном кластере. Кластеры объединяют, группируя объекты каждый раз во все более и более крупные кластеры. Этот процесс продолжают до тех пор, пока все объекты не станут членами одного единственного кластера.

Разделяющая, или дивизивная, кластеризация (divisive clustering) начинается со всех объектов, сгруппированных в единственном кластере. Кластеры делят (расщепляют) до тех пор, пока каждый объект не окажется в отдельном кластере.

Обычно в маркетинговых исследованиях используют агломеративные методы, например методы связи, дисперсионные и центроидные методы. Методы связи (linkage methods) включают метод одиночной связи, метод полной связи и метод средней связи.

В основе **метода одиночной связи** (single method) лежит минимальное расстояние, или правило ближайшего соседа. При формировании кластера первыми объединяют два объекта, расстояние между которыми минимально. Далее определяют следующее по величине самое короткое расстояние, и в кластер с первыми двумя объектами вводят третий объект. На каждой стадии расстояние между двумя кластерами представляет собой расстояние между их ближайшими точками.

**Метод полных связей.** Суть данного метода в том, что два объекта, принадлежащих одной и той же группе (кластеру), имеют коэффициент сходства, который меньше некоторого порогового значения . В терминах евклидова расстояния это означает, что расстояние между двумя точками (объектами) кластера не должно превышать некоторого порогового значения . Таким образом, определяет максимально допустимый диаметр подмножества, образующего кластер.

Метод **средней связи** (average linkage) действует аналогично. Однако в этом методе расстояние между двумя кластерами определяют как среднее значение всех расстояний, измеренных между объектами двух кластеров, при этом в каждую пару входят объекты из разных кластеров.

**Метод Варда.** В этом методе в качестве целевой функции применяют внутригрупповую сумму квадратов отклонений, которая есть ни что иное, как сумма квадратов расстояний между каждой точкой (объектом) и средней по кластеру, содержащему этот объект. На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров.

В **центроидных** методах (centroid method) расстояние между двумя кластерами представляет собой расстояние между их центроидами (средними для всех переменных). **Центроидный метод (centroid method)** Дисперсионный метод иерархической кластеризации, в котором расстояние между двумя кластерами представляет собой расстояние между их центроидами (средними для всех переменных). Каждый раз объекты группируют и вычисляют новый центроид. Изо всех иерархических методов методы средней связи и Варда показывают наилучшие результаты по сравнению с другими методами.

К другому типу процедур кластеризации относятся **неиерахические методы кластеризации** (nonhierarchical clustering), часто называемые методом -средних. Эти методы включают последовательный пороговый метод, параллельный пороговый метод и оптимизирующее распределение. В последовательном пороговом методе (sequential threshold method) выбирают центр кластера и все объекты, находящиеся в пределах заданного от центра порогового значения, группируют вместе. Затем выбирают новый кластерный центр, и процесс повторяют для несгруппированных точек. После того как объект помещен в кластер с этим новым центром, его уже не рассматривают как объект для дальнейшей кластеризации.

Аналогично работает **параллельный пороговый метод** (parallel threshold method), за исключением того, что одновременно выбирают несколько кластерных центров и объекты в пределах порогового уровня группируют с ближайшим центром.

Метод **оптимизирующего распределения** (optimizing partitioning method) отличается от двух изложенных выше пороговых методов тем, что объекты можно впоследствии поставить в соответствие другим кластерам, чтобы оптимизировать суммарный критерий, такой как среднее внутри кластерное расстояние для данного числа кластеров.

Недостатки неиерархических методов состоят в том, что число кластеров определяется заранее и выбор кластерных центров происходит независимо. Результаты кластеризации могут зависеть от выбранных центров. Многие неиерархические процедуры выбирают первые случаев *( –* число кластеров), не пропуская никаких значений в качестве начальных кластерных центров. Таким образом, результаты кластеризации зависят от порядка наблюдений в данных.

Неиерархическая кластеризация быстрее иерархических методов, и ее выгодно использовать при большом числе объектов или наблюдений.

Возможно использовать иерархические и неиерархические методы в тандеме. Во-первых, первоначальное решение по кластеризации получают, используя такие иерархические методы, как метод средней связи или метод Варда. Полученное этими методами число кластеров и кластерных центроидов используют в качестве исходных данных в методе оптимизирующего распределения.

**Данные.** Кластерный анализ можно применять к интервальным данным, частотам, бинарными данным. Важно, чтобы переменные изменялись в сравнимых шкалах.

Неоднородность единиц измерения и вытекающая отсюда невозможность обоснованного выражения значений различных показателей в одном масштабе приводит к тому, что величина расстояний между точками, отражающими положение объектов в пространстве их свойств, оказывается зависящей от произвольно избираемого масштаба. Чтобы устранить неоднородность измерения исходных данных, все их значения предварительно нормируются, т.е. выражаются через отношение этих значений  к некоторой величине, отражающей определенные свойства данного показателя. Нормирование исходных данных для кластерного анализа иногда проводится посредством деления исходных величин на среднеквадратичное отклонение соответствующих показателей.

Другой способ сводиться к вычислению, так называемого, стандартизованного вклада. Его еще называют -вкладом. -вклад показывает, сколько стандартных отклонений отделяет данное наблюдение от среднего значения:

,

где – значение данного наблюдения, – среднее, – стандартное отклонение. Среднее для -вкладов является нулевым и стандартное отклонение равно 1.

Стандартизация позволяет сравнивать наблюдения из различных распределений. Если распределение переменной является нормальным (или близким к нормальному), и средняя и дисперсия известны или оцениваются по большим выборным, то -вклад для наблюдения обеспечивает более специфическую информацию о его расположении.

Заметим, что методы нормирования означают признание всех признаков равноценными с точки зрения выяснения сходства рассматриваемых объектов. Уже отмечалось, что применительно к экономике признание равноценности различных показателей оправдано не всегда. Желательно наряду с нормированием придать каждому из показателей вес, отражающий его значимость в ходе установления сходств и различий объектов.

Экспертные оценки дают известное основание для определения важности индикаторов, входящих в ту или иную группу показателей. Умножение нормированных значений показателей на коэффициент, соответствующий среднему баллу оценки, позволяет рассчитывать расстояния между точками с учетом неодинакового веса их признаков.

Довольно часто при решении  подобных задач используют не один, а два расчета: первый, в котором все признаки считаются равнозначными, второй, где им придаются различные веса в соответствии со средними значениями экспертных оценок.

**Принятие решения о количестве кластеров.** Главный вопрос кластерного анализа – вопрос о количестве кластеров. Здесь нет твердых правил, позволяющих быстро принять решение, но можно руководствоваться следующим.

При определении количества кластеров руководствуются теоретическими и практическими соображениями. Например, если цель кластеризации – выявление сегментов рынка, то менеджмент может захотеть получить конкретное число кластеров.

В иерархической кластеризации в качестве критерия можно использовать расстояния, при которых объединяют кластеры.

В неиерархической кластеризации чертят график зависимости отношения суммарной внутригрупповой дисперсии к межгрупповой дисперсии от числа кластеров. Точка, в которой наблюдается изгиб или резкий поворот, указывает на приемлемое количество кластеров. Увеличение числа кластеров за эту точку обычно безрезультативно.

Относительные размеры кластеров должны быть достаточно выразительными.