Раздел 3. Методы анализа данных в социологии и маркетинге.

Тема 3.3. Кластерный (таксономический) анализ

Тема 3.4. Дискриминантный анализ.

1. Сущность кластерного анализа.
2. Статистики связанные с кластерным анализом.
3. Дискриминантный анализ**.**

Основная цель

Термин *кластерный анализ* (впервые ввел Tryon, 1939) в действительности включает в себя набор различных алгоритмов классификации. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как *организовать* наблюдаемые данные в наглядные структуры, т.е. развернуть таксономии. Например, биологи ставят цель разбить животных на различные виды, чтобы содержательно описать различия между ними. В соответствии с современной системой, принятой в биологии, человек принадлежит к приматам, млекопитающим, амниотам, позвоночным и животным. Заметьте, что в этой классификации, чем выше уровень агрегации, тем меньше сходства между членами в соответствующем классе. Человек имеет больше сходства с другими приматами (т.е. с обезьянами), чем с "отдаленными" членами семейства млекопитающих (например, собаками) и т.д.

Проверка статистической значимости

Заметим, что предыдущие рассуждения ссылаются на алгоритмы кластеризации, но ничего не упоминают о проверке статистической значимости. Фактически, кластерный анализ является не столько обычным статистическим методом, сколько "набором" различных алгоритмов "распределения объектов по кластерам". Существует точка зрения, что в отличие от многих других статистических процедур, методы кластерного анализа используются в большинстве случаев тогда, когда вы не имеете каких-либо априорных гипотез относительно классов, но все еще находитесь в описательной стадии исследования. Следует понимать, что кластерный анализ определяет "наиболее возможно значимое решение". Поэтому проверка статистической значимости в действительности здесь неприменима,

даже в случаях, когда известны p-уровни (как, например, в методе K средних).

Области применения

Техника кластеризации применяется в самых разнообразных областях. Хартиган (Hartigan, 1975) дал прекрасный обзор многих опубликованных исследований, содержащих результаты, полученные методами кластерного анализа. Например, в области медицины кластеризация заболеваний, лечения заболеваний или симптомов заболеваний приводит к широко используемым таксономиям. В области психиатрии правильная диагностика кластеров симптомов, таких как паранойя, шизофрения и т.д., является решающей для успешной терапии. В археологии с помощью кластерного анализа исследователи пытаются установить таксономии каменных орудий, похоронных объектов и т.д. Известны широкие применения кластерного анализа в маркетинговых исследованиях. В общем, всякий раз, когда необходимо классифицировать "горы" информации к пригодным для дальнейшей обработки группам, кластерный анализ оказывается весьма полезным и эффективным.

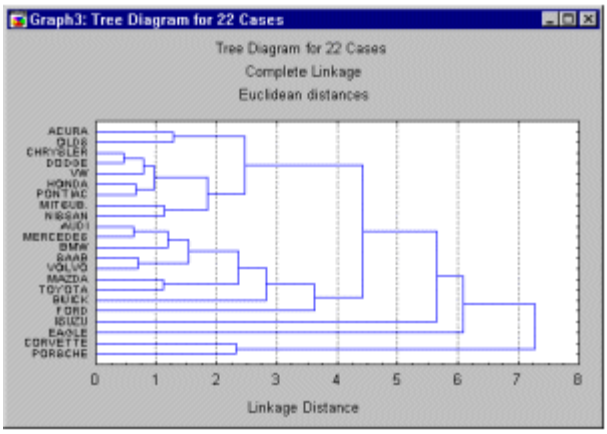
Объединение (древовидная кластеризация)

Общая логика

Приведенный ранее пример поясняет цель алгоритма объединения (*древовидной кластеризации*). Назначение этого алгоритма состоит в объединении объектов (например, животных) в достаточно большие кластеры, используя некоторую меру сходства или расстояние между объектами. Типичным результатом такой кластеризации является иерархическое дерево.

Иерархическое дерево

Рассмотрим *горизонтальную древовидную диаграмму*. Диаграмма начинается с каждого объекта в классе (в левой части диаграммы). Теперь представим себе, что постепенно (очень малыми шагами) вы "ослабляете" ваш критерий о том, какие объекты являются уникальными, а какие нет. Другими словами, вы понижаете порог, относящийся к решению об объединении двух или более объектов в один кластер.



В результате, вы *связываете* вместе всё большее и большее число объектов и агрегируете (*объединяете*) все больше и больше кластеров, состоящих из все сильнее различающихся элементов. Окончательно, на последнем шаге все объекты объединяются вместе. На этих диаграммах горизонтальные оси представляют расстояние объединения (в *вертикальных древовидных диаграммах* вертикальные оси представляют расстояние объединения). Так, для каждого узла в графе (там, где формируется новый кластер) вы можете видеть величину расстояния, для которого соответствующие элементы связываются в новый единственный кластер. Когда данные имеют ясную "структуру" в терминах кластеров объектов, сходных между собой, тогда эта структура, скорее всего, должна быть отражена в иерархическом дереве различными ветвями. В результате успешного анализа методом объединения появляется возможность обнаружить кластеры (ветви) и интерпретировать их.

Меры расстояния

Объединение или метод древовидной кластеризации используется при формировании кластеров несходства или расстояния между объектами. Эти расстояния могут определяться в одномерном или многомерном пространстве. Например, если вы должны кластеризовать типы еды в кафе, то можете принять во внимание количество содержащихся в ней калорий, цену, субъективную оценку вкуса и т.д. Наиболее прямой путь вычисления расстояний между объектами в многомерном пространстве состоит в вычислении евклидовых расстояний. Если вы имеете двух- или трёхмерное пространство, то эта мера является реальным геометрическим расстоянием между объектами в пространстве (как будто расстояния между объектами измерены рулеткой). Однако алгоритм объединения не "заботится" о том, являются ли "предоставленные" для этого расстояния настоящими или некоторыми другими производными мерами расстояния, что более значимо

для исследователя; и задачей исследователей является подобрать правильный метод для специфических применений.

Евклидово расстояние. Это, по-видимому, наиболее общий тип расстояния. Оно попросту является геометрическим расстоянием в многомерном пространстве и вычисляется следующим образом:

$$\text{расстояние}(x,y) = \{ \Sigma_i (x_i - y_i)^2 \}^{1/2}$$

Заметим, что евклидово расстояние (и его квадрат) вычисляется по исходным, а не по стандартизованным данным. Это обычный способ его вычисления, который имеет определенные преимущества (например, расстояние между двумя объектами не изменяется при введении в анализ нового объекта, который может оказаться выбросом). Тем не менее, на расстояния могут сильно влиять различия между осями, по координатам которых вычисляются эти расстояния. К примеру, если одна из осей измерена в сантиметрах, а вы потом переведете ее в миллиметры (умножая значения на 10), то окончательное евклидово расстояние (или квадрат евклидова расстояния), вычисляемое по координатам, сильно изменится, и, как следствие, результаты кластерного анализа могут сильно отличаться от предыдущих.

Квадрат евклидова расстояния. Иногда может возникнуть желание возвести в квадрат стандартное евклидово расстояние, чтобы придать большие веса более отдаленным друг от друга объектам. Это расстояние вычисляется следующим образом (см. также замечания в предыдущем пункте):

$$\text{расстояние}(x,y) = \Sigma_i (x_i - y_i)^2$$

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 расстояния Евклида. Однако отметим, что для этой меры влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат). Манхэттенское расстояние вычисляется по формуле:

$$\text{расстояние}(x,y) = \Sigma_i |x_i - y_i|$$

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:

расстояние(x,y) = Максимум$|x_i - y_i|$

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:

расстояние(x,y) = $(\Sigma_i |x_i - y_i|^p)^{1/r}$

где *r* и *p* - параметры, определяемые пользователем. Несколько примеров вычислений могут показать, как "работает" эта мера. 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 расстоянием Евклида.

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

расстояние(x,y) = (Количество $x_i \neq y_i$)/ i

Вопросы домашнего занятия:

1. В чем заключается сущность кластерного анализа?
2. Какие статистики связаны с кластерным анализом?
3. Назовите основные этапы выполнения кластерного анализа.
4. В каких случаях применяется неиерахическая кластеризация?
5. Как осуществляется кластеризация переменных?

Дискриминантный анализ .

Дискриминантный анализ используется для принятия решения о том, какие переменные различают (дискриминируют) две или более возникающие совокупности (группы). Например, некий исследователь в области

образования может захотеть исследовать, какие переменные относят выпускника средней школы к одной из трех категорий: (1) поступающий в колледж, (2) поступающий в профессиональную школу или (3) отказывающийся от дальнейшего образования или профессиональной подготовки. Для этой цели исследователь может собрать данные о различных переменных, связанных с учащимися школы. После выпуска большинство учащихся естественно должно попасть в одну из названных категорий. Затем можно использовать *Дискриминантный анализ* для определения того, какие переменные дают наилучшее предсказание выбора учащимися дальнейшего пути.

Медик может регистрировать различные переменные, относящиеся к состоянию больного, чтобы выяснить, какие переменные лучше предсказывают, что пациент, вероятно, выздоровел полностью (группа 1), частично (группа 2) или совсем не выздоровел (группа 3). Биолог может записать различные характеристики сходных типов (групп) цветов, чтобы затем провести анализ дискриминантной функции, наилучшим образом разделяющей типы или группы.

Вычислительный подход

С вычислительной точки зрения дискриминантный анализ очень похож на дисперсионный анализ. Рассмотрим следующий простой пример. Предположим, что вы измеряете рост в случайной выборке из 50 мужчин и 50 женщин. Женщины в среднем не так высоки, как мужчины, и эта разница должна найти отражение для каждой группы средних (для переменной *Рост*). Поэтому переменная *Рост* позволяет вам провести дискриминацию между мужчинами и женщинами лучше, чем, например, вероятность, выраженная следующими словами: "Если человек большой, то это, скорее всего, мужчина, а если маленький, то это вероятно женщина".

Вы можете обобщить все эти доводы на менее "тривиальные" группы и переменные. Например, предположим, что вы имеете две совокупности выпускников средней школы - тех, кто выбрал поступление в колледж, и тех, кто не собирается это делать. Вы можете собрать данные о намерениях учащихся продолжить образование в колледже за год до выпуска. Если средние для двух совокупностей (тех, кто в настоящее время собирается продолжить образование, и тех, кто отказывается) различны, то вы можете сказать, что намерение поступить в колледж, как это установлено за год до выпуска, позволяет разделить учащихся на тех, кто собирается и кто не собирается поступать в колледж (и эта информация может быть использована

членами школьного совета для подходящего руководства соответствующими студентами).

В завершение заметим, что основная идея дискриминантного анализа заключается в том, чтобы определить, отличаются ли совокупности по среднему какой-либо переменной (или линейной комбинации переменных), и затем использовать эту переменную, чтобы предсказать для новых членов их принадлежность к той или иной группе.

Дисперсионный анализ. Поставленная таким образом задача о дискриминантной функции может быть перефразирована как задача одновходового дисперсионного анализа (ANOVA). Можно спросить, в частности, являются ли две или более совокупности *значимо отличающимися* одна от другой по среднему значению какой-либо конкретной переменной. Однако должно быть ясно, что если среднее значение определенной переменной значимо различно для двух совокупностей, то вы можете сказать, что переменная разделяет данные совокупности.

В случае одной переменной окончательный критерий значимости того, разделяет переменная две совокупности или нет, дает *F*-критерий.

Многомерные переменные. При применении дискриминантного анализа обычно имеются несколько переменных, и задача состоит в том, чтобы установить, какие из переменных вносят свой вклад в дискриминацию между совокупностями. В этом случае вы имеете матрицу общих дисперсий и ковариаций, а также матрицы внутригрупповых дисперсий и ковариаций. Вы можете сравнить эти две матрицы с помощью многомерного *F*-критерия для того, чтобы определить, имеются ли значимые различия между группами (с точки зрения всех переменных). Эта процедура идентична процедуре *Многомерного дисперсионного анализа (MANOVA)*. Так же как в *MANOVA*, вначале можно выполнить многомерный критерий, и затем, в случае статистической значимости, посмотреть, какие из переменных имеют значимо различные средние для каждой из совокупностей. Поэтому, несмотря на то, что вычисления для нескольких переменных более сложны, применимо основное правило, заключающееся в том, что если вы производите дискриминацию между совокупностями, то должно быть заметно различие между средними.

Пошаговый дискриминантный анализ

Вероятно, наиболее общим применением дискриминантного анализа является включение в исследование многих переменных с целью определения тех из них, которые наилучшим образом разделяют совокупности между собой. Например, исследователь в области образования, интересующийся предсказанием выбора, который сделают выпускники средней школы относительно своего дальнейшего образования, произведет с целью получения наиболее точных прогнозов регистрацию возможно большего количества параметров обучающихся, например, мотивацию, академическую успеваемость и т.д.

Модель. Другими словами, вы хотите построить "модель", позволяющую лучше всего предсказать, к какой совокупности будет принадлежать тот или иной образец. В следующем рассуждении термин "в модели" будет использоваться для того, чтобы обозначать переменные, используемые в предсказании принадлежности к совокупности; о неиспользуемых для этого переменных будем говорить, что они "вне модели".

Пошаговый анализ с включением. В пошаговом анализе дискриминантных функций модель дискриминации строится по шагам. Точнее, на каждом шаге просматриваются все переменные и находится та из них, которая вносит наибольший вклад в различие между совокупностями. Эта переменная должна быть включена в модель на данном шаге, и происходит переход к следующему шагу.

Пошаговый анализ с исключением. Можно также двигаться в обратном направлении, в этом случае все переменные будут сначала включены в модель, а затем на каждом шаге будут устраняться переменные, вносящие малый вклад в предсказания. Тогда в качестве результата успешного анализа можно сохранить только "важные" переменные в модели, то есть те переменные, чей вклад в дискриминацию больше остальных.

***F* для включения, *F* для исключения.** Эта пошаговая процедура "руководствуется" соответствующим значением *F* для включения и соответствующим значением *F* для исключения. Значение *F* статистики для переменной указывает на ее статистическую значимость при дискриминации между совокупностями, то есть, она является мерой вклада переменной в предсказание членства в совокупности. Если вы знакомы с пошаговой процедурой множественной регрессии, то вы можете интерпретировать значение *F* для включения/исключения в том же самом смысле, что и в пошаговой регрессии.

Расчет на случай. Пошаговый дискриминантный анализ основан на использовании статистического уровня значимости. Поэтому по своей природе пошаговые процедуры рассчитывают на случай, так как они "тщательно перебирают" переменные, которые должны быть включены в модель для получения максимальной дискриминации. При использовании пошагового метода исследователь должен осознавать, что используемый при этом уровень значимости не отражает истинного значения *альфа*, то есть, вероятности ошибочного отклонения гипотезы H_0 (нулевой гипотезы, заключающейся в том, что между совокупностями нет различия).

Интерпретация функции дискриминации для двух групп

Для двух групп дискриминантный анализ может рассматриваться также как процедура множественной регрессии (и аналогичная ей); дискриминантный анализ для двух групп также называется *Линейным дискриминантным анализом Фишера* после работы Фишера (Fisher, 1936). (С вычислительной точки зрения все эти подходы аналогичны). Если вы кодируете две группы как *1* и *2*, и затем используете эти переменные в качестве зависимых переменных в множественной регрессии, то получите результаты, аналогичные тем, которые получили бы с помощью *Дискриминантного анализ*а. В общем, в случае двух совокупностей вы подгоняете линейное уравнение следующего типа:

$$\text{Группа} = a + b_1*x_1 + b_2*x_2 + ... + b_m*x_m$$

где a является константой, и $b_1...b_m$ являются коэффициентами регрессии. Интерпретация результатов задачи с двумя совокупностями тесно следует логике применения множественной регрессии: переменные с наибольшими регрессионными коэффициентами вносят наибольший вклад в дискриминацию.

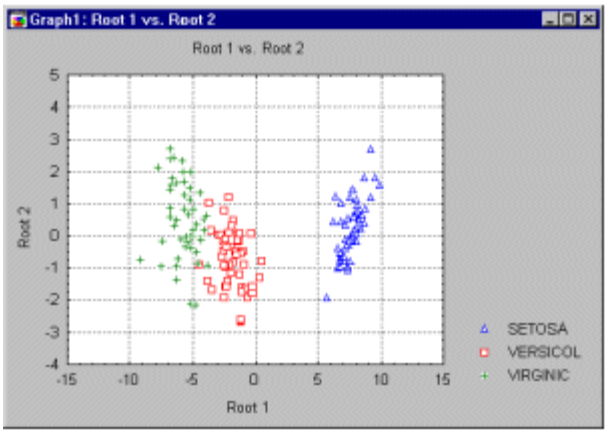
Дискриминантные функции для нескольких групп

Если имеется более двух групп, то можно оценить более, чем одну дискриминантную функцию подобно тому, как это было сделано ранее. Например, когда имеются три совокупности, вы можете оценить: (1) - функцию для дискриминации между совокупностью 1 и совокупностями 2 и 3, взятыми вместе, и (2) - другую функцию для дискриминации между совокупностью 2 и совокупности 3. Например, вы можете иметь одну функцию, дискриминирующую между теми выпускниками средней школы, которые идут в колледж, против тех, кто этого не делает (но хочет получить

работу или пойти в училище), и вторую функцию для дискриминации между теми выпускниками, которые хотят получить работу против тех, кто хочет пойти в училище. Коэффициенты *b* в этих дискриминирующих функциях могут быть проинтерпретированы тем же способом, что и ранее.

Канонический анализ. Когда проводится дискриминантный анализ нескольких групп, вы не должны указывать, каким образом следует комбинировать группы для формирования различных дискриминирующих функций. Вместо этого, вы можете автоматически определить некоторые оптимальные комбинации переменных, так что первая функция проведет наилучшую дискриминацию между всеми группами, вторая функция будет второй наилучшей и т.д. Более того, функции будут независимыми или *ортогональными*, то есть их вклады в разделение совокупностей не будут перекрываться. С вычислительной точки зрения система вы проводите анализ *канонических корреляций*, которые будут определять последовательные канонические *корни* и функции. Максимальное число функций будет равно числу совокупностей минус один или числу переменных в анализе в зависимости от того, какое из этих чисел меньше.

Интерпретация дискриминантных функций. Как было установлено ранее, вы получите коэффициенты *b* (и стандартизованные коэффициенты *бета*) для каждой переменной и для каждой дискриминантной (теперь называемой также и *канонической*) функции. Они могут быть также проинтерпретированы обычным образом: чем больше стандартизованный коэффициент, тем больше вклад соответствующей переменной в дискриминацию совокупностей. (Отметим также, что вы можете также проинтерпретировать *структурные коэффициенты*; см. ниже.) Однако эти коэффициенты не дают информации о том, между какими совокупностями дискриминируют соответствующие функции. Вы можете определить характер дискриминации для каждой дискриминантной (канонической) функции, взглянув на средние функций для всех совокупностей. Вы также можете посмотреть, как две функции дискриминируют между группами, построив значения, которые принимают обе дискриминантные функции (см., например, следующий график).



В этом примере *Корень1 (root1*), похоже, в основном дискриминирует между группой *Setosa* и объединением групп *Virginic* и *Versicol*. По вертикальной оси (*Корень2*) заметно небольшое смещение точек группы *Versicol* вниз относительно центральной линии (*0*).

Матрица факторной структуры. Другим способом определения того, какие переменные "маркируют" или определяют отдельную дискриминантную функцию, является использование факторной структуры. Коэффициенты факторной структуры являются корреляциями между переменными в модели и дискриминирующей функцией. Если вы знакомы с факторным анализом, то можете рассматривать эти корреляции как факторные *нагрузки* переменных на каждую дискриминантную функцию.

Некоторые авторы согласны с тем, что структурные коэффициенты могут быть использованы при интерпретации реального "смысла" дискриминирующей функции. Объяснения, даваемые этими авторами, заключаются в том, что: (1) - вероятно структура коэффициентов более устойчива и (2) - они позволяют интерпретировать факторы (дискриминирующие функции) таким же образом, как и в факторном анализе. Однако последующие исследования с использованием метода Монте-Карло (Барсиковский и Стивенс (Barcikowski, Stevens, 1975); Хьюберти (Huberty, 1975)) показали, что коэффициенты дискриминантных функций и структурные коэффициенты почти одинаково нестабильны, пока значение размер выборки не станет достаточно большим (например, если число наблюдений в 20 раз больше, чем число переменных). Важно помнить, что коэффициенты дискриминантной функции отражают уникальный (частный) вклад каждой переменной в отдельную дискриминантную функцию, в то время как структурные коэффициенты отражают простую корреляцию между переменными и функциями. Если дискриминирующей функции хотят придать отдельные "осмысленные" значения (родственные

интерпретации факторов в факторном анализе), то следует использовать (интерпретировать) структурные коэффициенты. Если же хотят определить вклад, который вносит каждая переменная в дискриминантную функцию, то используют коэффициенты (веса) дискриминантной функции.

Значимость дискриминантной функции. Можно проверить число корней, которое добавляется *значимо* к дискриминации между совокупностями. Для интерпретации могут быть использованы только те из них, которые будут признаны статистически значимыми. Остальные функции (корни) должны быть проигнорированы.

Итог. Итак, при интерпретации дискриминантной функции для нескольких совокупностей и нескольких переменных, вначале хотят проверить значимость различных функций и в дальнейшем использовать только значимые функции. Затем, для каждой значащей функции вы должны рассмотреть для каждой переменной стандартизованные коэффициенты *бета*. Чем больше стандартизованный коэффициент *бета,* тем большим является относительный собственный вклад переменной в дискриминацию, выполняемую соответствующей дискриминантной функцией. В порядке получения отдельных "осмысленных" значений дискриминирующих функций можно также исследовать матрицу факторной структуры с корреляциями между переменными и дискриминирующей функцией. В заключение, вы должны посмотреть на средние для значимых дискриминирующих функций для того, чтобы определить, какие функции и между какими совокупностями проводят дискриминацию.

Вопросы домашнего задания:

1. Каковы особенности проведения дискриминантного анализа?
2. Как дискриминантный анализ связан с регрессионным и дисперсионным анализом?
3. В чем состоит модель дискриминантного анализа?
4. Укажите этапы проведения дискриминантного анализа?
5. Назовите особенности каждого этапа.
6. Какие статистики связаны с дискриминантным анализом?
7. Каковы особенности проведения множественного дискриминантного анализа?
8. Каковы особенности проведения пошагового дискриминантного анализа?