Кластерный анализ

Структура лекции

1. Многомерный анализ
2. Кластерный анализ

Многомерный анализ

При исследовании объекта измеряется сразу несколько характеристик.

Часть математической статистики, которая исследует эксперименты с такими многомерными наблюдениями, называется *многомерным статистическим анализом.*

Хорошо разработана математическая теория для многомерных гауссовских наблюдений, т.е. для случайных величин, подчиняющихся многомерному нормальному распределению. Здесь почти для каждого одномерного гауссовского метода существует соответствующий многомерный вариант.

Имеются решения и для некоторых специфически многомерных статистических проблем.

Факторный анализ

Факторный анализ дает возможность количественно определить нечто, непосредственно неизмеряемое, исходя из нескольких доступных измерению переменных. Факторный анализ позволяет установить для большого числа исходных признаков сравнительно узкий набор «свойств», характеризующих связь между группами этих признаков и называемых факторами.

Процедура факторного анализа состоит из четырех основных стадий.

1. Вычисление корреляционной матрицы для всех переменных, участвующих в анализе.
2. Извлечение факторов.
3. Вращение факторов для создания упрощенной структуры.
4. Интерпретация факторов.

Извлечение факторов

Извлечение фактора начинается с подсчета суммарного разброса значений всех участвующих в анализе переменных (данная величина чем-то похожа на общую сумму квадратов).

Первой задачей факторного анализа является выбор взаимодействующих переменных, чья взаимная корреляция обусловливает наибольшую долю общей дисперсии. Эти переменные образуют *первый фактор.*

Затем первый фактор исключается и из оставшегося множества переменных снова выбираются те, чье взаимодействие определяет наибольшую долю оставшейся общей дисперсии. Эти переменные образуют *второй фактор.* Процедура извлечения факторов продолжается до тех пор, пока не будет исчерпана вся общая дисперсия переменных.

Выбор и вращение факторов

Для исследователя не представляют интереса все извлеченные факторы. Если факторов окажется столько же, сколько исходных переменных, факторный анализ теряет смысл, поскольку его целью является сокращение исходного набора переменных.

Нужно принять решение, какие из факторов следует оставить для дальнейшего анализа.

В первую очередь, рекомендуется руководствоваться здравым смыслом и оставлять те факторы, которые имеют понятную теоретическую или логическую интерпретацию.

Не всегда представляется возможным заранее установить назначение каждого фактора, и поэтому исследователи на первом этапе обычно используют формальные критерии. Однако обычно факторы, полученные методом главных компонент, не поддаются достаточно наглядной интерпретации. Поэтому следующим шагом факторного анализа служит преобразование (вращение) факторов таким образом, чтобы облегчить их интерпретацию.

Дискриминантный анализ

Предположим, что мы имеем совокупность объектов, разбитую на несколько групп (т.е. для каждого объекта можно сказать, к какой группе он относится). Пусть для каждого объекта имеются измерения нескольких количественных характеристик.

Нужно найти способ, как на основании этих характеристик можно узнать группу, к которой принадлежит объект. Это позволит для новых объектов из той же совокупности предсказывать группы, к которой они относятся.

Для решения этой задачи применяются методы *дискриминантного анализа,* они позволяют строить функции измеряемых характеристик, значения которых и объясняют разбиение объектов на группы.

Многомерное шкалирование

Во многих областях исследования (например, в психологии, биологии, социологии, лингвистике и т.д.) бывает затруднительно или невозможно проводить непосредственное измерение интересующих исследователя характеристик объектов из изучаемой совокупности, зато можно экспертным или каким-то другим путем оценивать степень сходства или различия между парами объектов.

В этом случае для интерпретации получаемых данных используются методы многомерного шкалирования.

Они позволяют представить совокупность интересующих исследователя объектов в виде некоторого набора точек многомерного пространства некоторой небольшой размерности, при этом каждому объекту соответствует одна точка. Координаты точек истолковываются как значения неких характеристик исходных объектов, которые и объясняют их свойства или взаимоотношения.

В случае удачного шкалирования, когда точки полученного пространства представляют объекты без серьезных погрешностей и размерность этого пространства невелика (равна, двум или трем), исследователь получает возможность представить изучаемую совокупность объектов наглядно. Часто это помогает по-новому осознать проблему, увидеть ее новые черты и особенности либо осознать те скрытые признаки, которые и определяют видимые свойства объектов или их взаимоотношения.

Часто в качестве исходных данных для шкалирования используются не сами оценки степени сходства объектов, а результаты их ранжирования. Соответствующие методы шкалирования называются *неметрическими.*

*И*сходными данными часто служат суждения человека (как испытуемого либо как эксперта), поэтому их количественные значения носят в значительной мере условный характер. Чтобы избавиться от этой условности, и прибегают к ранжированию.

Кластерный анализ

Методы кластерного анализа позволяют разбить изучаемую совокупность объектов на группы «схожих» объектов, называемых *кластерами.*

Большинство методов кластеризации (иерархической группировки) являются *агломеративными* (объединительными)

Графическое изображение процесса объединения кластеров может быть получено с помощью *дендрограммы* — дерева объединения кластеров.

Другие методы кластерного анализа являются *дивизивными* — они пытаются разбивать объекты на кластеры непосредственно.

Методы кластеризации довольно разнообразны, в них по-разному выбирается способ определения близости между кластерами (и между объектами), а также используются различные алгоритмы вычислений.

Результаты кластеризации зависят от выбранного метода, и эта зависимость тем сильнее, чем менее явно изучаемая совокупность разделяется на группы объектов. Поэтому результаты вычислительной кластеризации могут быть дискуссионными.

Методы кластерного анализа не дают какого-либо способа для проверки статистической гипотезы об адекватности полученных классификаций. Иногда результаты кластеризации можно обосновать с помощью методов дискриминантного анализа.

В пакете SPSS представлены все перечисленные выше методы. При этом есть возможность гибкого выбора и настройки параметров соответствующих процедур. Например, иерархическая кластеризация в пакете предусматривает задание различных расстояний между объектами, различных методов объединения объектов в кластеры и т. п.

В документации и во встроенном справочнике SPSS можно найти дополнительные пояснения по назначению и методике применения статистических методов.

Кластерный анализ

Программа SPSS реализует три метода кластерного анализа:

- Двухэтапный кластерный анализ (TwoStep),
- Кластеризация К-средними (K-means),
- Иерархическая кластеризация (Hierarchical)

Двухэтапный кластерный анализ позволяет выявить группы (кластеры) объектов по заданным переменным, если эти группы действительно существуют. При этом программа автоматически определяет количество существующих кластеров (групп). Если невозможно однозначно определить количество кластеров, все объекты помещаются в один.

Кластеризация К-средними разбивает по заданным переменным все множество объектов на заданное пользователем число кластеров так, чтобы средние значения для кластеров по каждой из переменных максимально различались.

Иерархическая кластеризация, как наиболее гибкий из рассматриваемых методов, позволяет детально исследовать структуру различий между объектами и выбрать наиболее оптимальное число кластеров. В силу этого иерархический кластерный анализ применяется наиболее часто.

Этапы кластерного анализа

Кластерный анализ выполняется в несколько этапов, приводящих к конечному результату.

Выделяют несколько этапов кластерного анализа.

1. *Выбор переменных-критериев для кластеризации.*
2. *Выбор способа измерения расстояния между объектами, или кластерами* (изначально считается, что каждый объект соответствует одному кластеру). По умолчанию используется квадрат Евклидова расстояния, согласно которому расстояние между объектами равно сумме квадратов разностей между значениями одноименных переменных объектов.

3. Формирование кластеров. Существует два основных метода формирования кластеров: *метод слияния* и *метод дробления.* В первом случае исходные кластеры увеличиваются путем объединения до тех пор, пока не будет сформирован единственный кластер, содержащий все данные. Метод дробления основан на обратной операции; сначала все данные объединяются в один кластер, который затем делится на части до тех пор, пока не будет достигнут желаемый результат. По умолчанию программой SPSS используется метод слияния.

4. *Интерпретация результатов.* Как и в случае факторного анализа, желаемое число кластеров и оценка результатов анализа зависят от целей исследователя