А.Б. Домрачева

Лекции по математическому моделированию

2019 г.

Содержание

1 Характеристики инженерных задач

Прикладные задачи классифицируют как экономические, инженерные и научные. Отличительные особенности инженерных задач:

1. инженерные задачи имеют ярко-выраженную практическую направленность (создание новых конструкций, разработка техпроцесса, минимизация затрат на производство изделий и т.д.);

2. для инженерных задач характерна необходимость доведения результатов до конкретных чисел, на основании которых можно принимать решения;

3. для инженерных задач характерен значительный объем выполняемой вычислительной работы, при этом набор методов и их реализаций конечен;

4. для инженерных задач характерна необходимость создания достаточно сложных математических моделей с использованием современных вычислительных методов. Экономические задачи используют простые типовые модели. Научные задачи, наоборот, выходят за рамки сложных, но типовых моделей и требуют разработки или модификации математического аппарата для их решения и реализации;

5. инженерные задачи решаются, как правило, специалистами в предметном области (не программистами, не мат-моделистами, не прикладными математиками).

> «Нельзя научить делать открытия, но можно подготовить открытие»
>
> А.Б. Домрачева

За это отвечает теория инженерного эксперимента — один из разделов математического моделирования. В свою очередь, теория инженерного эксперимента делится на теории планирования эксперимента и проведения эксперимента и обработку результатов эксперимента.

1.1 Проблемы, возникающие при решении инженерных задач

Проблема первая. Характеристики объектов испытаний, которые требуется определить в результате эксперимента, часто оказываются недоступными непосредственному измерению.

Проблема вторая. Процессы функционирования объекта исследования часто имеют сложный динамический характер и подвержены существенным влияниям изменяющихся условий внешней среды.

Проблема третья. При испытании сложных комплексов необходимо учитывать влияние испытательного, регистрирующего, управляющего оборудования (погрешность измерения).

Таким образом, эксперимент может оказаться продолжительным и дорогостоящим. В связи с этим необходим системный подход, предполагающий решение описанных выше проблем с обеспечением необходимой точности решения.

Математическое моделирование — метод исследования объектов и процессов реального мира с помощью их приближенных описаний под средством формул (равенств, неравенств, уравнений, логических структур). Такое описание называют математическими моделями.

Математическая модель — неотъемлемая часть эксперимента, исследуемая и уже известная техниками (?) экономической характеристики объекта.

Выделяют класс плохо-формализуемых задач, для которого известны собственные методы построения моделей.

Процесс создания математической модели:

1. построение математической модели;
2. постановка исследования и решение вычислительных задач, реализующих модель;
3. проверка качества моделей на практике, модификация моделей.

1.2 Построение математической модели

Имея результаты эксперимента, можно установить связь между этими величинами, описываемую на языке математики. При выборе известной математической модели (модель должна быть достаточно полной для изучения свойств исследования или объекта и достаточно простой для реализации на ЭВМ) выбор учитываемых существенных и несущественных факторов носит исследовательский характер: как правило подтверждается характеристиками. В противном случае компромисс матмодели будет определен неверно.

Прежде чем использовать ту или иную известную матмодель необходимо определить тип модели: аналитическая, гипотетическая, имитационная, статическая или динамическая.

1. гипотетическая модель — еще не подтвержденный процесс;
2. статическая модель — описывает явления в предположении, что процесс завершен;
3. динамическая модель — описывает явления или процессы реального мира при переходе от одного состояния к другому;
4. имитационная модель — позволяет имитировать поведение реальной системы на ЦВМ в заданный или формируемый период времени с определенным набором входных данных;

5. аналитическая модель — формализация явления или процесса реального мира, учитывающая физико-химические воздействия, протекающие в системе.

Пример: баллистическая задача. Пусть тело (шар) брошено под углом α к поверхности Земли со скоростью v_0. Пренебрегая сопротивлением воздуха, считая Землю плоской, а $g = \text{const}$, определить дальность полета.

Модель Галилея.

картинка

$$\begin{cases} x = v_0 \cdot \cos\alpha \cdot t, \\ y = v_0 \cdot \sin\alpha \cdot t - g\frac{t^2}{2}. \end{cases}$$

$$\begin{cases} x_1 = 0, \\ y_1 = 0 \end{cases}, \begin{cases} x_2 = \operatorname{tg}\alpha \frac{2v_o \cos^2\alpha}{g}, \\ y_2 = \frac{-g}{2v_0^2 \cos^2\alpha} x^2 + (\operatorname{tg}\alpha)x \end{cases}$$

Модель Ньютона: постановка задачи совпадает, но учитывается сила лобового сопротивления воздуха $F = -\beta v^2$, где $\beta = \frac{c\rho S}{2}$ — коэффициент сопротивления воздуха.

картинка

$$\begin{cases} m\frac{du}{dt} = -\beta u\sqrt{u^2 + v^2}, \\ m\frac{dw}{dt} = -\beta w\sqrt{u^2 + v^2} - mg. \end{cases}$$

Модель Ньютона более точная в сравнении с моделью Галилея, но не удовлетворяет решению современной баллистической задачи (линейные и угловые параметры, всего шесть неизвестных).

Имея результаты эксперимента (выраженные в числовых значениях), необходимо установить связь между этими величинами, описываемую формулам. Модель будут определять величины, которые можно назвать:

1. исходными данными;
2. параметрами модели;
3. выходными данными.

В моделировании, как правило, модель представляется в виде «черного ящика», входные данные — вектор $\bar{X}$, выходные данные — вектор $\bar{Y}$, параметры — A.

картинка

В соответствии со схемой выделяют типы решаемых задач:

1. по входным данным и набору параметров при известном наборе преобразований входных данных в выходные требуется найти решение;
2. по значениям выходных данных $\bar{Y}$ и фиксированных значениях параметров A (при известном наборе преобразований входных данных в выходные) необходимо оценить набор исходных данных $\bar{X}$;

В модели Галилея решение обратной задачи требует минимум двух переменных в векторе $\bar{Y}$.

Ранее как обратные классифицировались и задачи идентификации в узком смысле, то есть такие задачи, в которых известны наборы $\bar{X}$ и $\bar{Y}$, а также тип связи, и достаточно было оценить параметры.

Задача идентификации считается заданной в широком смысле, когда не имеется никаких сведений о типе преобразования (ранее считались не имеющими решения). Пример: в СЛАУ $A\bar{x} = \bar{b}$ найти A при известных $\bar{x}, \bar{b}$.

Даже для простых моделей реализация использует два и более вычислительных алгоритма, то есть после формализации модели на первом этапе необходимо проверить корректность постановки каждой задачи по Адамару-Петровскому и выбрать вычислительный метод решения каждой задачи, обеспечив хорошую обусловленность как самой задачи, так и алгоритма.

После первого расчета по выбранной модели выясняют пригодность для описания процесса или объекта реального мира.

Теоретические выводы и результаты, полученные из математической модели, сопоставляются с экспериментами. Может выясниться, что модель адекватна, то есть вполне точно описывает объект исследования, иначе же требуются модификации. На этом этапе решение принимают специалисты в предметной области, оценивая полученную точность и рентабельность модификаций модели.

Для проведения эксперимента:

1. составляется план;
2. создается экспериментальная установка;
3. выполняются контрольные эксперименты;
4. проводятся серийные опыты;

5. обрабатываются полученные экспериментальные данные и их результаты.

Натурные эксперименты как правило дороги, часто продолжительны. Дешевой и быстродейственной альтернативой натурному эксперименту является *вычислительный эксперимент*, в основе которого лежит реализация математическом модели на ЦВМ. Возникает коллизия: для построения модели необходимо провести натурный эксперимент (что бывает невозможно). С другой стороны матмодель строится как альтернатива натурному эксперименту. Эта коллизия считалась нерушимой до середины двадцатого века.

Компромиссом является так называемый *полунатурный эксперимент*, где в качестве модели используются физическая модель объекта, доступные характеристики оцениваются экспериментально (н: число Маха в аэродинамической трубе), другие — программно.

Преимущества вычислительного эксперимента:

1. дешевле;
2. безопасен;
3. можно повторить;
4. предполагает моделирование условий, которые нельзя создать в лаборатории.

Ограничения:

1. применимость результата ограничена рамками используемой математической модели.

Создание нового экспериментального цикла предусматривает анализ альтернатив и вариантов эксперимента, а также оптимизацию по заданному критерию или ряду параметров. Для получения оптимального результата может понадобиться большое количество вариантов оптимизации, что делает заведомо продолжительными и вычислительный эксперимент. Для сокращения времени эксперимента строят план вычислительного эксперимента, для чего в настоящее время существует ПО, позволяющее автоматизировать процесс планирования и проведения эксперимента (т.н. проблемно-ориентированное ПО), что вместе с вычислительной техникой и обслуживающим персоналом составляет проблемно-ориентированную информационную систему исследований.

2 Основные этапы решения задачи с применением ЦВМ

Решение инженерной задачи разделяется на 10 этапов:

1. постановка задачи;
2. построение или модификация математической модели;
3. постановка вычислительной задачи;
4. предварительный предмашинный анализ свойств вычислительной задачи;
5. выбор и построение математической модели;
6. алгоритмизация и программирование;
7. отладка программы;
8. счет по программе;
9. обработка и интерпретация результатов;
10. использование результатов (в т.ч. для коррекции матмодели).

2.1 Постановка задачи

Имеется общая предварительная формулировка. Установка конкретизируется с уточнением цели исследования. Определяется система координат, в которой задаются входные и выходные данные, начальные и конечные условия. Находится компромисс между полезностью результатов и сложностью достижения цели. Задача формализуется на языке предметной области (но с учетом возможностей современной вычислительной техники).

2.2 Построение математической модели

Анализ объекта исследования. Меняется ли он динамически в заданный или формируемый период времени, носят ли переменные (входные или выходные) или алгоритм их преобразования случайный характер, либо нет. На основе анализа цели исследования оценивается необходимость решения прямой, обратной или идентификационной задачи.

1. анализируется возможность измерения известных согласно постановке задачи данных; строится схема объекта исследования в виде черного ящика с учетом количества компонент и изменения поведения системы во времени;

2. выбирается тип моделируемой системы (простая, структурно-сложная, сложная с изменяющимся поведениям, гибридная, ...). На схеме указываются все измеряемые переменные и параметры, указывается способ их получения;

3. определяется тип базовой вычислительной задачи (прямая, обратная, идентификационная);

4. указывается классифицирующие признаки модели (статическая или динамическая, аналитическая или имитационная, вероятностная или статистическая, ...);

5. по результатам классификации с учетом опыта предметной области выбирается, модифицируется или строится матмодель.

В случае необходимости построения матмодели учитывается опыт смежных областей, а также поиск аналогов.

При построении модели разного типа имеются принципиальные различия, в связи с чем используются типовые методики, в том числе *методика динамической модели* действий или процесса.

Методика динамической модели:

1. рисуется графическая схема, иллюстрирующая процесс. Например: в выбранной системе координат указываются время начала и, по возможности, время конца измерений. Альтернативой является блок-схема процесса;

2. по схеме уточняются и обозначаются начальные и конечные входные данные, выходные данные полученные в процессе моделирования, выделяются переменные, которые являются выходными на одном этапе решения задачи и выходными на другом;

3. определяются физико-химические воздействия на объект, влияющие на изменения выходных данных;

4. оценивается возможность упрощения модели, ряд воздействий не учитывается;

5. формализуется общее уравнение модели;

6. конкретизируется описание каждого отдельного воздействия.

Любая модель предполагает формулировку ограничений, накладываемых на неё. Эти ограничения указываются на этапе постановки задачи и на графической схеме могут обозначаться как вычеркивания физико-химических воздействий на систему.

На этом этапе уточняется цель исследования и на основе цели формализуются требования к точности вычислений результата и к точности представлений исходных данных. В ряде случаев исходные данные также являются результатом расчетов (если не заданы, не могут быть измерены

на основе натурного эксперимента, не выражаются через другие известные либо измеряемые характеристики аналитически). Натурный эксперимент чаще всего является дорогостоящим, объект исследования в нем — непосредственен. Объект исследования задачи, в ряде случаев — продолжительный. При этом построение математической модели на основе вычислительного эксперимента приводит к основной коллизии математического моделирования. В этом случае проводят полунатурный эксперимент, объектом исследования которого является физическая модель объекта исследования. Результат, полученный с помощью полунатурного эксперимента может оказаться недостаточно точным. В связи с чем появляется дополнительная вычислительная задача, итерационно повышающая точность оценки. В ряде случаев уточнения требуют и результат решения всей задачи, то есть в рамках моделирования формализуется основная вычислительная задача, и ряд вспомогательных, определяющих входные (выходные) данные и параметры модели. Проводится классификация задачи (прямая, обратная, идентификационная), в ряде случаев основная задача также разбивается на несколько вычислений, которые решить проще.

картинка

Каждая полученная задача анализируется на корректность постановки по Адамару-Петровскому. Если какая-либо из задач является некорректно поставленной, то ее необходимо переформулировать с целью обеспечения существования и единственности решения, оценить обусловленность задачи, анализировать возможность обеспечения устойчивости и хорошей обусловленности решения. Можно переходить к выбору и построению численных методов и последующей алгоритмизации. Перед выбором численного метода задачи пытаются решить количественно, то есть получить количественный результат, но с невысокой точностью. При этом становится известным диапазон значений для каждого решения и вырабатываются ограничения, накладываемые на данные или метод, в том числе подтверждается возможность решения задачи . Чаще решение инженерной задачи сводится к последовательному решению стандарных вычислительных задач для который разработаны эффективные численные методы, но возможны ситуации, когда алгоритм необходимо адаптировать к решению задачи или создать новый метод.

Если имеется несколько методов решения, необходимо сравнить эффективность всей совокупности методов. Для сравнения эффективности выбирают показатели по которым будут сравнивать методы и составляют таблицу эффективности: по горизонтали — название методов, по вертикали — значения показателей этого метода.

2.3 Алгоритмизация и программирование, отладка и счет по программе

Численный метод содержит только принципиальную схему решения, в связи с чем на этом этапе разрабатывается вычислительный алгоритм. Проверяется корректность, обеспечивается устойчивость решения, в том числе вычислительная, результат должен достигаться за конечное число операций.

При тестировании алгоритма осуществляется не только поиск синтаксических ошибок, но и проводится валидация результата на тестовых данных. Если тестовые данные не могут быть заданы произвольно, то предварительно решается задача, обратная поставленной. Результат выводится в удобной форме: графики, таблицы, диаграммы на основании многократного запуска с неизменными и изменяемыми входными данными. Многократный запуск с неизменными данными подтверждает устойчивость и хорошую обусловленность решения, изменение результата оценивается с точки зрения допустимой погрешности, в случае отрицательного результата необходимо вернуться на этап анализа вычислительной задачи. При многократном запуске с изменяемыми данными подтверждается обусловленность задачи, устойчивость решения, проверяются ограничения, наложенные при постановке задачи.

2.4 Обработка и интерпретация результатов

На заключительном этапе оценивается соответствие результатов вычислительного эксперимента (в основе лежит построенная матмодель) явлениям или процессам реального мира. Для подтверждения приходится проводить новые натурные испытания, но, в данном случае, коллизия не наблюдается:

1. модель построена по результатам другого полунатурного эксперимента;
2. как правило, выходной натурный эксперимент не так трудоемок, в ряде случаев требуется два-три контрольных значения показателей исследуемого объекта.

Следует отметить, что вычислительный эксперимент с целью подтверждения адекватности модели может проводиться по уникальному плану в отличие от типовой процедуры валидации и верификации результатов расчета на этапе отладки и счета по программе.

3 Статические модели

Статические модели активно применяются при сравнении «срезов», полученных в разные моменты времени той или иной динамической модели. Статически приближают: поверхность объекта, а также изменяемые географические (с изменением координат) показатели исследуемых систем. Как правило, статические модели используются в случае сложности динамических, а также для наглядоности представления.

Статическая модель может оказаться лишь иллюстративной частью проекта, использующего физическую модель исследуемого объекта или его динамическую математическую модель.

3.1 Методика построения статической модели

1. рисуется графическая схема, иллюстрирующая состояние системы в некоторый момент времени в заданной системе координат;

2. по схеме уточняются входные данные (н: матрица высот), уточняется погрешность представления измеряемой координаты. Выделяются промежуточные переменные, по которым будут оценены выходные, и непосредственно выходные данные.

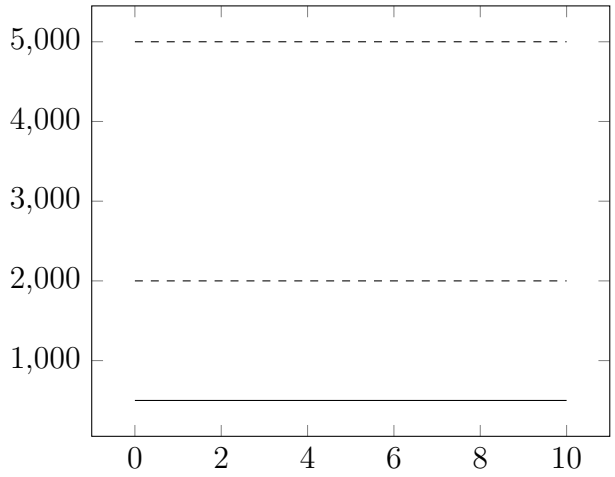

Рис. 1:

   В отличие от динамической модели, параметры статической на начальном этапе моделирования не так очевидны, и могут быть оценены по результатам вспомогательных задач.

3. определяются физико-химические воздействия на объект, переводящие в данный момент систему из предыдущего в текущее состояние;

4. оценивается возможность формализации модели и её упрощения. Необходимо обосновать возможность описания и целесообразность учета выбранных на предыдущем этапе физико-

химических воздействий на исследуемый объект. Если в динамических моделях большее значение имеет учет физических воздействий, то в статических — химических воздействий;

5. формулируется на языке математики общее уравнение, описывающее состояние объекта в заданный период времени, с учетом выделенных физико-химических воздействий;

6. конкретизируется описание состояния объекта, однозначо выделяются входные, выходные, промежуточные переменные, а также параметры модели;

7. при формализации модели должно выполняться условие единственности решения, в силу чего ряд входных данных может быть доопределен, либо исключен.

Если динамическая модель сложна, то часто выделяют состояние системы, вносящее значительный вклад в решение задачи. В этом случае динамическую систему моделируют с помощью конечного числа статических систем.

Динамическое моделирование ресурсозатратно, но дает более точное представление о развитии процесса реального мира. Статическое моделирование описывает процесс дискретно, то есть не дает полной картины его развития. При этом демострирует качественное (а не количественное) решение при низкой ресурсозатратности.

4 Классификация динамических систем

В системе могут изменяться два и более объектов, кроме того возможно взаимодействие (обмен данными) между объектами, что говорит о сложных структуре и поведении динамических систем.

Различают простые динамические системы, структурно-сложные динамические системы, сложные динамические системы, меняющие поведение во времени, структурно-сложные гибридные системы.

4.1 Простые динамические системы

Определение: обычно понимается система, поведение которой задается совопкупностью обыкновенных дифференциальных уравнений в форме Коши с достаточно гладкими правыми частями, обеспечивающими существование и единственность решения. Примерами объектов являются: современная баллистическая задача, задача с бассейном (в одну трубу вливается, в другую выливается) и др.

В ряде случаев система может быть представлена в виде нескольких простых динамических систем. Тогда такая система тоже считается просто динамической.

4.2 Структурно-сложные динамические системы

Большинство технических и природных систем являются более сложными, при наличии нескольких взаимодействующих компонент системы говорят о структрурно-сложной динамической системе.

Уже отмечалось, что структура современных моделей соответствует структуре изучаемого объекта, но функционирование каждой компоненты системы скрыто от наблюдателя. Известны только входные (в ряде случаев выходные) данные и структура системы. Модель представляется в виде черного ящика с указанием количества компонентов и схемы их взаимодействия:

картинка (черный ящик с четыремя компонентами; $\bar{X}$, $\bar{Y}$)

При моделировании работы каждой компоненты можно получать сведения об изменении их состояния в т.ч. с учетом взаимодействий. При параллельном функционировании компонент в системе требуется синхронизация такого взаимодействия.

4.3 Сложные динамические системы, меняющие поведение во времени

Одной из черт сложного поведения системы является наличие у системы нескольких сменяющих друг друга состояний. Такие изменения можно представить блочной стуктурой и описывать как простую динамическую систему с определенными параметрами нагружения на разных уровнях. В случае периодической нагрузки это оправдано, но в случае стохастического изменения поведения описание оказывается сложным.

картинка с уровнями (ступеньками)

Непрерывное изменение поведения системы можно дискретизировать, понимая, что на границе блоков необходимо исследовать переходную функцию. В противном случае модель будет иметь значительную погрешность.

Погрешность моделирования — осознанное пренебрежение при построении матмодели. Далее обозначется как δ_{mod}.

Причинами разрывов функций может быть использование недостаточно гладких функий для описания дискретного блока. Такие системы называются *гибридными*, и помимо описания разных стилей поведения системы требует исследования функций, формализующих это поведение.

4.4 Структурно-сложная гибридная система

Соединяет черты структурно-сложных и гибридных систем. Сложность моделирования делает почти невозможной использование аналитических моделей в силу чего такой класс систем предполагает построение имитационных моделей.

5 Имитационное моделирование

Рассматриваемые выше статические и динамические модели описывают однокомпонентные системы и относятся к классу аналитических, учитывают физико-химические воздействия на систему. Если система много-компонента, то есть описывается взаимодействием ее составных частей (объектов), аналитические модели оказываются слишком трудоемкими и ресурсозатратными при их реализации на ЦВМ. Альтернативой является имитационное моделирование, не учитывающее конкретных физико-химических процессов, протекающих в системе, а лишь приближающее закон изменения входных данных при получении определенного набора выходных данных.

В основе идеи имитационного моделирования лежит идея «черного ящика». То есть имитационное моделирование сводится к решению задач идентификации. Представление оператора A унифицируется.

На систему может действовать набор управляющих воздействий (вектор U), под действием которых динамически меняется набор параметров системы. Необходимость управляющих воздействий и их изменений возникает в связи с несоблюдением ограничений, наложенных как на входные так и на выходные данные. Таким образом, должны быть сформулированы ограничения для входных и выходных данных.

На рисунке представлен так называемый агрегат, соответствующий одной компоненте в системе. А компонент может быть несколько. А их функционирование как последовательным, так и параллельным или последовательно-параллельным.

Под эмуляцией понимают точное повторение функциональных возможностей системы с использованием иного алгоритма преобразования входных данных в выходные. Имитация с некоторой точностью воспроизводит функциональные возможности системы, применяя иной алгоритм преобразования входных данных в выходные (аппроксимационный метод). Симуляция - настройка параметров системы на основе ряда входных данных с целью получения хотя бы грубого приближения.

Таким образом, симуляционным моделированием обычно называют стохастическую настройку системы с целью получения приближенного (иногда только качественного) результата. В ряде задач большего и не требуется. В случае имитационного моделирования параметры системы рассчитываются, с целью получения выходных данных с незначительной погрешностью.

5.1 Модельное время

При функционировании объекта реального мира в течение заданного или формируемого периода времени говорят о реальном времени. Время, затраченное на программную реализацию модели функционирующего объекта, называют вычислительным временем. Реальное время измеряется в единицах, указанных при постановке задачи. Вычислительное время, как правило, на несколько порядков меньше и измеряется в секундах или долях секунды. Модельное время измеряется в

единицах реального времени. При этом оно дискретно: каждый квант модельного времени соответствует переходу системы из одного состояния в другое.

Модельное время — это дискретное время, измеряемое в единицах реального времени, связанное с вычислительным временем и используемое для внешней и внутренней синхронизации модели.

Очевидна невозможность последовательного моделирования компонент системы, а параллельная имитация не соответствует структуре задачи, в силу чего принято использовать следующие схемы организации квазипараллелизма (псевдопараллелизма):

1. активностями;
2. событиями;
3. транзактами;
4. процессами;
5. агрегатами.

В этом случае модельное время оказывается связанным с вычислительным временем, что способствует внутренней и внешней синхронизации модели. При этом делает возможным моделирование взаимодействия компонент.

6 Преобразование формального описания в имитационную модель

Функции, как правило, возложены на управляющую программу моделирования (УПМ).

6.1 Внутренняя синхронизация компонент модели

Переход от формального описания к имитационной модели осуществляется на основе декомпозиции сложной системы на составные части. Для каждой компоненты кроме её формального описания устанавливают временную координату. Таким образом, УПМ реализует операторы коррекции модельного времени.

Если требуется дополнительная декомпозиция компонент модели (активностей, событий, процессов, агрегатов) и шага модельного времени, то может возникнуть необходимость в декомпозиции формального описания алгоритма.

Для транзактного способа имитации процедура декомпозиции не требуется, так как просто меняется количество заявок и интенсивность их поступления.

Агрегатный подход по организации внутренней синхронизации похож на транзактный, с той разницей, что вместо задержки, имитирующей выполнение программы, реализуется тот или иной вычислительный алогоритм.

При моделировании событиями любое дополнительное событие включается в расписание в виде времени возникновения события.

При моделировании процессами и активностями внутренняя синхронизация обеспечивается с помощью «семафоров» — глобальных переменных, значения которых либо разрешают, либо запрещают запуск программы-имитатора. В случае запрета активизации все компоненты модели помещатся в очередь. Далее исполняются по очереди в соответствии с заданными приоритетами.

Для транзактного способа имитации средством внутренней синхронизации компонент являются сами транзактные очереди.

При агрегатном способе дополнительно огранизуются (программно) управляющие сигналы.

Остальные подходы требуют не только активирующих сигналов, но и, в ряде случаев, контроля за ходом имитации, проверок окончания счета, контроля за взаимодействием компонентов.

6.2 Внешняя синхронизация компонент модели

Используются операторы синхронизации компонент в системе дискретных событий. При детальном описании компонент реализуется оператор внезапного перехода в ждущий режим или оператор перевода в ждущий режим до выполнения условия. Такими операторами функционирование модели делится на части, в рамках которых УПМ может запустить иной алгоритм описания компонент.

Для транзактного и агрегатного способа внешняя синхронизация не требуется в силу унификации решаемых задач.

6.3 Синхронизация моделей управления информацией

Важно, чтобы результаты функционирования одной компоненты или процесса появились не раньше времени τ_{ij}, и были использованы другой компонентой не позже времени $\tau_{ij} + \Delta\tau_{ij}$. В связи с этим используют глобальную переменную, которая активирует процесс записи-считывания данных в информационном поле.

При имитации активностями и событиями дополнительная процедура не требуется. В остальных случаях необходимо осуществлять обращение к УПМ.

6.4 Решение конфликтных ситуаций в модели

Аппарат внешней и внутренней синхронизации, а также синхронизация моделей управления информацией с помощью операторов останова и запуска системы разрешает конфликты на основе системы приоритетов.

Особое место занимают системы, в которых компоненты динамически меняют своё поведение. Как правило составляются иерархические структуры, позволяющие УПМ найти оптимальный маршрут, разрешив при этом конфликтную ситуацию. В этом случае дополнительно требуются операторы, инициирующие активности, события, процессы, а также деинициализирущие их.

6.5 Организация контроля за ходом имитации

Помимо определенных задачей остановок или запусков системы возможны энергосбои и пр. Необходимо провести журнализацию состояния модели, то есть сохранить значения, определяющие модель, состояние оперативной памяти и т.д. При восстановлении условий моделирования работа будет продолжена без потерь (включая состояние УПМ).

При транзактном способе оперативный контроль ??? блоком сбора статистики. При агрегатном — по аналогии с транзактным. Прочие методы огранизуются просто, и могут быть созданы программистом в рамке работы над УПМ.

6.6 Огранизация сбора статистики

Устанавливаются так называемые «наблюдатели-модели»: используются все моменты времени, выбранные установленной моделью для принудительного окончания счета, для обмена данными. Это функция сбора статистики, например:

1. число взаимодействий компонент модели в некоторый период времени;

2. среднее время пребывания заявки в системе;

3. количество потерянных заявок;

4. ...

Большинство функций реализовано аналитически.

Принято создавать наблюдателя как единицу модели, например наблюдатель-активность или наблюдатель-процесс. Как правило, функция автоматически реализована в системе, но может быть расширена программистом.

6.7 Окончание имитации

Аналогично реализуется блок окончания моделирования. Он может отражать последовательность действий по контролю за моментам окончания имитации и начала обработки результатов моделирования. Кроме того, реализует планирование очередного имитационного эксперимента.

6.8 Документирование

Протоколирует результат моделирования: описание имитационной модели, входных, промежуточных и прочих данных.

7 Транзактный способ организации псевдопараллелизма

Общем особенностью всех транзактных трансляторов являются понятия: *транзакция* (заявка), *устройство*, *канал* (по которому заявки поступают в систему) и *очередь*, а также некоторое количество единиц памяти в общем накопителе.

7.1 Оператор генерации транзактов GPSS

`GENERATE 12, 4, 50, 5, 1` — будут сгенерированы случайные величины по равномерному закону распределения в диапозоне: []. Первый транзакт появится с задержкой 50 единиц модельного времени. Всего будет создано 5 транзактов. Приоритет транзактов равен единице. Если 5 не написать (но запятую оставить), то количество транзактов не ограничено.

`GENERATE 12, FN$FFF, 50, 5, 1` — интеравал времени между транзактами есть целая часть числа 12 умножить на `$FFF`.

`FNK FUNCTION, R, N1, C4 0.0/0.1, 0.8/0.5, 1.6/1, 0, 1.9` — описание функции, аргуметом которой является случайная величина, равномерно распределенная в интрвале $[0, 1]$. Функция задана таблично четыремя точками.

`SEIZE PLOT` — занятие устройства `PLOT` транзактом.

`RELEASE PLOT` — освобождение устройства `PLOT` уже обслуженным транзактом.

`ADVANCE A, B` — задержка транзакта на время, длительность которого определяют парамеры: $t \in [A - B, A + B]$.

`QUEUE SQV`— оператор огранизации очереди (`SQV` — это имя). При появлении каждого транзакта длина очереди увеличивается на единицу.

`DEPART SQV` — освободить одно место в очереди.

`ENTER MEM, 12` — занятие транзактом двенадцати единиц памяти в накомителе (`MEM` — это тоже название).

`LEAVE MEM, *2` — освобождает несколько единиц памяти в накопителе (количество единиц хранится во втором параметре транзакта).

`TRANSFER, MET` — безусловный переход по метке `MET`.

`TRANSFER BOTH, L1, L2` — переход по метке `L1` если это возможно, иначе в `L2`. Если и это невозможно, то транзакт задерживается до следующего момента дискретного модельного времени.

8 Имитационное моделирование

В отличие от процесса моделирования статических или простых динамических систем, в имитационном моделировании участвует несколько специалистов (а, в ряде случаев, структурных подразделений). Различают квалификации специалистов предметной области, математиков и IT-специалистов.

Пунткы методики:

1. постановка задачи в терминах предметной области, классификация сложной динамической системы (структурно-сложная, меняющая поведение во времени, гибридная и др.);
2. построение структурной схемы исследуемой системы с целью выделения компонент и определения взаимодействия между ними;
3. по схеме уточняются известные входные данные, параметры системы, определяются неизвестные промежуточные переменные, выходные данные, определяются ограничения, обратная (положительная или отрицательная) связь, управляющие воздействия на систему, начальные и краевые условия задачи. На этом этапе модель окончательно классифицируется как сложная, предварительно выбирается схема организации квазипараллелизма модели;
4. вместе со специалистом в предметной области строится концептуальная модель исследуемой системы;
5. преобразование концептуальной модели в имитационную;
6. уточнение значений входных переменных и параметров системы или получение методики оценивания этих значений;
7. формализация модели;
8. отладка модели на ЦВМ;
9. счет по программе;
10. проверка качества модели на практике, оценка необходимости модификации модели;
11. совместно со специалистами предметной области и IT-специалистами выявляются «узкие места» модели.

    Для сокращения обсуждения уточнение модели проводится начиная с этапа концептуального проектирования. В редких случаях дефект может быть уже на этапе описания предметной области (первый пункт).

12. проведени вычислительного эксперимента с использованием построенной математической модели.

8.1 Преобразование концептуальной модели в имитационную

На входе имеется предварительная схема, построенная на основе описания предметной области. На этом этапе уточняется набор входных и выходных данных и параметров системы, методики их определения, ограничения, обратная связь (при наличии) и воздействия окружающей среды или управляющия воздействия.

На первом этапе задача была поставлена в терминах предметной области без учета возможностей вычислительной техники и современного математического аппарата. Учитывая современные возможности реализации, ряд компонент могут быть объединены либо, напротив, декомпозированы. Таким образом, структура модели может значительно отличаться от структуры исследуемой системы, что, как правило, упрощает решение задачи, снижая его точность.

Система представляется в виде набора «черных ящиков», для каждого из которых **однозначно** определен набор входных и выходных данных и алгоритм преобрзования входных данных в выходные.

На основе уточненной схемы выбирается (уточняется) схехма организации псевдо-параллелизма, соответствующая реализуемости модели.

Итогом этапа является выбор схемы организации псевдо-параллелизма, и для *качественных решений*: событийный и транзактный подход, а для получения адекватных *количественных оценок*: моделирование активностями и агрегатами, с целью получения *прогноза* — как правило, процессный подход.

9 Основные понятия систем массового обслуживания

В общем случае постановка задачи теории СМО записывается так: в систему поступают заявки, сгенерированные по одному или разным законам распределения (известны заранее). Заявки обрабатываются втечение случайного времени (как правило, определяется интенсивностью обработки). Для обработки заявок выделяются пять секунд n устройств (или «каналов»). Необходимо по таким параметрам:

1. решить прямую задачу
   (a) среднее время нахождения заявки в очереди;
   (b) скорость обработки заявок в системе;
   (c) число потерянных заявок;
2. решить обратую задачу
   (a) определить число устройств, необходимое для обработки заявок с заданной вероятностью (обычно $\in [0.9, 1.0]$).

Различают системы с ограниченным размером очередии без ограничений. Принципиально случай ограничения очереди не отличается от неограниченной очереди за исключением учета потерянных заявок. Если для модели этот показатель принципиален, то организуют очередь с ограничением.

9.1 Задача анализа СМО

Анализ сводится к оцениванию показателей эффективной работы системы на основе которых могут быть приняты управленческие решения, обеспечивающие эффективную работу системы (н: увеличить количество столиков в кафе). Такие показатели делятся на четыре группы:

1. показатели, характеризиующие систему в целом: число обслуженных заявок, число занятых каналов, среднее время пребывания заявки в системе, количество заявок, ожидающих обслуживания, потерянных заявок;
2. вторичные вероятностные показатели, которые оцениваются на основе показателе первой группы: вероятность того, что заявка будет обслужена или потеряна, вероятность пустой очереди и прочее;
3. экономические показатели, характеризующие стоимость потерь, связанных с необслуженными заявками и наоборот, с дополнительным обслуживанием заявок и прочее;

4. прочие показатели, затраты ресурсов на обработку (потерю) заявок.

Показатели 1-3 унифицированы, включены в статистику инструментальных средств анализа СМО. Показатели четвертой группы могут быть рассчитаны программно.

9.2 Математическое описание СМО

СМО рассматриваются как некоторые физические системы с дискретными состояниями $S_0, S_1, \ldots, S_n$ функционирующие при том в реальном времени t. Число таких состояний может быть как конечно, так и бесконечно. Для анализа системы строится граф переходов, где вершинами являются состояния, а ребрам соответствуют вероятности переходов из одного состояния в другое. Переход может быть возвратным и невозвратным (одно- или двунаправленное ребро). По аналогии с графом составляется таблица переходов (матрица смежности).

Переход из состояния в состояние происходит в случайный момент времени t_{ij}. В ряде случаев переходы описывают в виде потоков: если заявки однотипны, то выделяют поток установления на обслуживание, поток прибывания в очереди, поток обслуживания заявок и поток освобождения очереди. То есть, математическое моделирование СМО сводится к описанию на основе Марковских Цепей (не принципиально каким образом заявка оказалась в состоянии S_i, но принципиально в какое следующее состояние и с какой вероятностью она перейдет).

ПРОПУЩЕНА ЛЕКЦИЯ

Рассмотрим СМО в некоторый момент времени t и, задав некоторый малый промежуток времени Δt, оценим вероятность того, что в момент времени $t + \Delta t$ система не выйдет из состояния S_0. Возможно два случая (???)

1. Пусть в момент времени t с некоторой вероятностью $p_0(t)$ система находилась в состоянии S_0. Возможно суммировать переход из S_0 в S_1 и из S_1 в S_2 или сразу перейти в состояние S_2. Вероятность того, что мы останемся в состоянии S_0 равна $p_{00} = 1 - (p_{01} + p_{02})$, где p_{ij} — вероятность перехода из состояния S_i в состояние S_j.:

   Индуктивно, $\lambda_0 = 1 - (\lambda_{01} + \lambda_{02})$ — интенсивность переходов, $p = \Delta t \lambda$. Итого, имеем:

$$\Delta t \lambda_0 = \Delta t \left[1 - (\lambda_{01} + \lambda_{02})\right],$$

$$\hat{p_{00}} = 1 - \Delta t(\lambda_{01} + \lambda_{02}).$$

2. Система с вероятностью $p_1(t)$ находится в состоянии S_1 или в состоянии S_2 с вероятностью $p_2(t)$. Необходимо оценить вероятность того, что за время Δt система вернется в состояние S_0.

$$p_0(t+\Delta t) = p_1(t)\cdot\lambda_{10}\Delta t + p_2(t)\cdot\lambda_{20}\Delta t + p_0(t)\cdot[1-(\lambda_{01}+\lambda_{02})\Delta t],$$
$$\frac{p_0(t+\Delta t)}{\Delta t} = -p_0(t) + p_1(t)\lambda_{10} + p_2(t)\lambda_{20} - (\lambda_{01}+\lambda_{02})p_0(t),$$
$$p_0'(t) \approx p_1(t)\lambda_{10} + p_2(t)\lambda_{20} - (\lambda_{01}+\lambda_{02})p_0(t),$$
$$p_1(t)\lambda_{10} + p_2(t)\lambda_{20} - (\lambda_{01}+\lambda_{02})p_0(t) \underset{\Delta t\to 0}{\to} p_0'(t).$$

Индуктивно расспространяя рассуждения на другие состояния системы, получим:

$$\begin{cases} p_0'(t) = p_1(t)\lambda_{10} + p_2(t)\lambda_{20} - (\lambda_{01}+\lambda_{02})p_0(t), \\ p_1'(t) = p_0(t)\lambda_{01} + p_3(t)\lambda_{31} - (\lambda_{10}+\lambda_{13})p_1(t), \\ p_2'(t) = p_0(t)\lambda_{02} + p_3(t)\lambda_{32} - (\lambda_{20}+\lambda_{23})p_2(t), \\ p_3'(t) = p_1(t)\lambda_{13} + p_2(t)\lambda_{23} - (\lambda_{31}+\lambda_{32})p_3(t). \end{cases}$$

Из этой системы следует, что вероятности $p_i(t) \to p_i$, $i = \overline{0,\dots,n}$. Причем эти вероятности с течением времени стремятся к некоторомым значениям p_i, называемым *предельной (финальной) вероятностью*. В ряде случаев требуется оценить именно финальные вероятности. Т.к. финальные вероятности константы, то

$$\begin{cases} \lambda_{10}p_1(t^*) + \lambda_{20}p_2(t^*) = (\lambda_{01}+\lambda_{02})p_0(t^*), \\ \lambda_{01}p_0(t^*) + \lambda_{31}p_3(t^*) = (\lambda_{01}+\lambda_{13})p_1(t^*), \\ \lambda_{02}p_0(t*) + \lambda_{32}p_3(t^*) = (\lambda_{20}+\lambda_{23})p_2(t^*), \\ \lambda_{13}p_1(t^*) + \lambda_{23}p_2(t^*) = (\lambda_{31}+\lambda_{32})p_3(t^*)). \end{cases}$$

При условии известных интенсивностей перехода и системы выше оцениваются величины предельных вероятностей.

Важно: выкладки выше приведены для конкретной системы (четыре состояния S_i), а не в общей форме!

▷ Пример. Четыре состояния S_0, S_1, S_2, S_3. Вероятности переходов:

$$\lambda_{01} = 1, \quad \lambda_{10} = 2,$$
$$\lambda_{02} = 2, \quad \lambda_{20} = 3,$$
$$\lambda_{13} = 2, \quad \lambda_{31} = 3,$$
$$\lambda_{23} = 2, \quad \lambda_{32} = 3.$$

В ряде случаев уравнения в системе могут оказаться линейно-зависимыми. В таких случаях систему определяют краевыми или начальными условиями:

$$\begin{cases} 3p_0 = 2p_1 + 3p_2, \\ 4p_1 = p_0 + 3p_3, \\ 4p_2 = 2p_0 + 2p_3, \\ \underline{p_1 + p_2 + p_3 = 1}. \end{cases} \Longrightarrow \begin{cases} \hat{p_0} = 0.10, \\ \hat{p_1} = 0.20, \\ \hat{p_2} = 0.27, \\ \hat{p_3} = 0.13. \end{cases}$$

«Найти средний чистый доход от эксплуатации в стационарном режиме исследуемой системы, если исправная работа первого и второго узла приносит доход 10 и 6 у/е соотвественно, а ремонт требует затрат 4 и 2 у/е соответственно. Оценить имеющуюся возможность уменьшения вдвое среднего времени ремонта каждого из двух узлов, если при этом вдвое увеличиваются затраты на ремонт каждого узла в единицу времени».

Решим задачу с использованием постренной системы. Примем обозначения: S_0 — база, S_1 — первый станок, S_2 — второй станок, S_3 — ремонт.

$$p_0 + p_3 = 0.4 + 0.2 = 0.6, \qquad \text{(I)}$$
$$p_0 + p_2 = 0.4 + 0.27 = 0.67, \qquad \text{(II)}$$
$$p_1 + p_3 = 0.2 + 0.13 = 0.33, \qquad \text{(I, ремонт)}$$
$$p_2 + p_3 = 0.27 + 0.13 = 0.4. \qquad \text{(II, ремонт)}$$

Итого, доход равен $0.6 \cdot 10 + 0.67 \cdot 6 - 0.33 \cdot 4 - 0.4 \cdot 2 = 7.9$ у/е. $\triangleleft$

10 Стохастическое моделирование

10.1 Основные задачи математической статистики и теории вероятности. Вероятностные и статистические модели

Мат. статистика (МС) и теория вероятностей (ТВ) изучает математические модели случайных явлений или процессов. Задачи МС являются обратными по отношению к задачам ТВ: в ТВ после задания некоторого случайного явления требуется рассчитать вероятностные характеристики этого явления или процесса. Приведенные выше примеры являются вероятностными моделями. В задачах МС известны результаты эксперимента, и требуется оценить неизвестные параметры модели.

Вторая задача — это задача *интервального* оценивания неизвестных параметров. При этом требуется построить интервал с неизвестными границами, в который анализируемый параметр попадет с заданной вероятностью.

Напоминание: доверительный интервал и коэффициент доверия:

$$P\left\{\underline{a} \leq a \leq \overline{a}\right\} = \gamma,$$

$\underline{a}, \overline{a}$ — нижняя и верхняя границы доверительного интервала;
γ — коэффициенет доверия. Обычно выбирается равным одному из значений: 0.9, 0.95, 0.99, 0.995, 0.999, 0.9995.

Третья задача МС — задача проверки статистических гипотез, в которой требуется на основе данных эксперимента проверить то или иное предположение.

Следует отличать статистические гипотезы от бытовых (н: «После обеда пойдет дождь»). Статистическая гипотеза является формализацией бытовой, и, формулируется относительно закона распределения параметров модели или, непосредственно, самих параметров модели.

В МС часто используется выборочная терминология, основанная не следующей «урновой» схеме: пусть имеется урна, содержащая N чисел $x_1, x_2, \ldots, x_N$, скрытых от наблюдателя. Весь этот набор называется *генеральной совокупностью*. Из генеральной совокупности случайным образом выбирается конечное число n значений. Такой набор значений $x_1, \ldots, x_n$ называется выборкой объема n из генеральной совокупности. Различают выборки с и без *возвращения* (обратно в генеральную совокупность). Если выборка производится с возвращением, то случайные величины независимы. В противном случае зависимы. В первом случае такой набор называется *независимой повторной случайной выборкой* объема n.

В случае бесконечной генеральной совокупности терминология сохраняется, но в этом случае разница между выборками с возвращением и без фактически изсчезает.

Если упорядочить выборку $x_1, \ldots, x_n$ в порядке неубывания (реже: невозрастания):

$$x^{(1)}, x^{(2)}, \ldots, x^{(n)},$$
$$x^{(1)} \leq x^{(2)} \leq \ldots \leq x^{(n)},$$

то такая выборка назыается *вариационным рядом*.

Важно: не путать с временным рядом!

Для описания выборки в целом и вариационного ряда в частности используют империческую функцию распределения.

Прежде чем дать общее описание, рассмотрим ▷ пример. Рассмотрим выборку $\{1025, 1031, 1048, 1092,$

$$\hat{F}_k(x) = \begin{cases} 0, & x \leq x_0, \\ \frac{j}{n}, & x_0 < x < x_{max}, \\ 1, & x \geq x_{max}, \end{cases}$$

где n — общее число точек в выборке, k — число разбиений, j — число значений, попавших в диапазон $[x_0, x_i]$ (i — порядковый номер интервала). Можно приблизить имперической функцией

$$\hat{F}_k(x) = \frac{r(x)}{n},$$

$r(x)$ — число возможных чисел в выборке, для которых $x \in [x_0, x_i]$.

По теореме Гливено-Кантелли при $n \to \infty$ с вероятностью 1 выполняется соотношение

$$\sup_x \left| \hat{F}_n(x) - G(x) \right| \to 0.$$

На основании теоремы Гливено-Кантелли можно сделать вывод, что эмперическая функция распределения сходится к теоретической. ◁

10.2 Моделирование случайных векторов и случайных процессов

В силу дискретизации данных на ЦВМ генерация случайного вектора и реализаций N на $[0, \tau]$ случайного процесса принципиально не отличаются друг от друга.

В отличие от независимой повторной выборки случайных величин, элементы вектора и реализации СП являются зависимыми.

10.2.1 Моделирование станционарных СП с распространенными одномерными законами распределения

Например, случайный процесс с показательным распределением может быть записан рекурентной формулой

$$\xi(t_n) = \rho_n \xi(t_{n-1}) + \sqrt{1 - \rho_n^2} x[n]$$

$$\rho_n = e^{-f(t_n - t_{n-1})}$$

$x[n]$ — последовательность независимых нормальных случайных величин с параматерами 0, 1. То есть необходимо сгенерировать N случайных величин, распределенных по стандартному нормальному закону, после чего обеспечить зависимость реализации $\xi(t_i)$, $i = \overline{1, N}$ на основе формулы (10.2.1). В силу того, что формула (10.2.1) получена на основе экспериментальных данных, полученная реализация оказывается очень неточным приближением случайного процесса с показательным определением. Кроме того, эмпирические формулы существуют не для всех типов распределения S_p случайного процесса, либо они вычислительно сложны.

рисунок с 3д колоколом

$$x_1, x_2, \ldots, x_N, \quad PP[0, 1].$$

$$\begin{cases} x_i^* = a_i + (b_i - a_i) x_i, \quad i = 1, N, \\ x_0^* = f_{max} \cdot x_0. \end{cases}$$

Если $f(x_1^*, x_2^*, \ldots, x_N^*) \leq x_0^*$, то вектор аргументов используется как реализация случайного вектора $\{x_1^*, \ldots, x_N^*\}$. В противом случае набор отбрасывается.

Основным недостатоком метода Неймана является искусственное ограничение области определения аргументов заданной плотности распределения. Для большинства распределений погрешность метода оказывается не столь существенной.

<КОНСПЕКТЫ НАТАШИ>

Такие задачи решаются на основе марковских цепей. Цепью Маркова с дискретным временем называется описание процесса, при котором вероятность перехода в $(n+1)$-ое состояние зависит от того, в каком состоянии находится система в текущий момент времени и не зависит от того, как система перешла в это состояние.

Округление ресурсного значения всегда происходит в большую сторону (н: патронов, снарядов должно быть в среднем 4 для полного поражения цели). Если вероятность перехода в новое состояние также не зависит от номера шага, говорят об *однородной марковской цепи в дискретном времени.*

Марковские цепи с непрерывным временем также предполагают некоторый набор состояний. Формула описания марковской цепи с непрерывным временем:

$$P(X_{t+h} = x_{t+h} \mid X_s = x_s, 0 < s \leq t) = P(X_{t+h} = x_{t+h} \mid X_t = x_t).$$

Цепь Маркова с непрерывным временем называется *однородной*, если

$$P(X_{t+h} = x_{t+h} \mid X_t = x_t) = P(X_h = x_h \mid X_0 = x_0).$$

Вероятность того, что система будет находиться в состоянии S_i и не выйдет из него за время Δt связано с интенсивностью потока событий и может быть оценена по формуле

$$P_0(t)\left[1 + (\lambda_{01} + \lambda_{02})\Delta t\right] = p_0(t) + p_0(t)\lambda\Delta_t, \lambda = \lambda_{01} + \lambda_{02}, (???)$$

где $P_0(t)$ — вероятность события остаться в том же состоянии, Δt — исследуемый период времени, λ_{01} — интенсивность переходов из S_0 в S_1, λ_{02} — интенсивность переходов из S_0 в S_2.

Построенные на основе имперической функции распределения моменты называют *выборочными* или *империческими.* $\hat{x} = \frac{1}{n}\sum x_i$ называется выборочным средним (империческим средним).

10.3 Аналог первого начального момента или мат. ожидания случайной величины x

Величина $\hat{S}^2 = \frac{1}{n}(\sum x_i\hat{x})^2$ называется *выборочной* или *имперической* дисперсией, является выборочным аналогом второго центрального момента (дисперсии) случайной величины x. В ряде случаев используются характеристики выборки, связанные с начальными или центральными моментами, при этом определяющие размах полученных величин. Примером такой характеристики является СКО. Для больших выборок с возможно повторяющимися значениями $\sum n_i = n$.

Пусть рассматриваются две выборки случайных величин x и y. Связь между показателями x и y может быть более или менее тесной. В случае достаточно представительных выборок эту связь можно описать функциональной зависимостью. Такую возможность нужно подтвердить, изучив корреляцию (в том числе взаимную корреляцию) двух выборок. Если связь оказывается недостаточно тесной, можно исследовать только стахостическую (случайную) зависимость. Кроме того,

выборки могут быть недостаточно представительны. Выборками среднего объема являются выборки 20-100 элементов, при $n \geq 100$ являются представительными. Выборки меньше 20 элементов являются непредставительными, в этом случае разрабатываются специальные методы постоения феноменологических моделей. Неудовлитворительные свойства оценок параметров функций связи между этими показателями.

Перечисленные проблемы не дают возможность однозначно моделировать объект или процесс реального мира и приводят к необходимости построения феноменологических моделей, построенных исключительно на результатах опыта.

10.4 Точечные оценки параметров

Пусть имеется некоторая случайная величина $x \in X$ с функцией распределния $F(x, \theta)$ и плотностью распределения $f(x, \theta)$. В общем случае параметр θ может быть многомерным. Параметр θ требуется оценить по серии испытаний наблюдения над величиной x на основе повторной независимой случайной выборки $x_1, x_2, \ldots, x_n$. Совокупность распределений $F(x_i, \theta)$, $i = 1, \ldots, n$ называется *параметрическим* семейством распределний. Любая функция ϕ, не зависящая от результатов эксперимента, называется *статистикой*. Тогда задача оценивания параметра θ сводится к нахождению такой функции от результата наблюдения (статистики), что оценка θ в некотором смысле близка к истинному значению параметра θ распределния $F(x, \theta)$.

10.5 Свойства оценок

Оценка $\hat{\theta}$ наызвается *несмещенной оценкой* параметра θ, если ее математическое ожидание равно самому этому показателю при любом возможном значении параметра θ. Оценка $\hat{\theta}$ называется *асимптотически несмещенной* оценкой параметра θ, если $\hat{\theta} \to \theta$. На свойства оценки влияет также объем выборки. Оценка $\hat{\theta_n}$ называется *состоятельной оценкой* параметра θ, если при $n \to \infty$ она сходится по вероятности к истинному значению параметров.

Пусть $\hat{\theta}_n$ — асимптотически несмещенная оценка параметра θ и $\hat{\theta}_n \to 0$, то оценка считается эффективной строго *по сравнению* с оценкой $\hat{\theta}_n^{(2)}$.

10.6 Точная оценка параметров метода моментов и метода максимального правдоподобия

Пусть $x_1, x_2, \ldots, x_n$ — независимая случайная повторная выборка из некоторой генеральной совокупности с плотностью распределения $f(x, \theta)$ (прим: θ — вектор в этой лекции) и $F(x, \theta)$. Необходимо оценить $\hat{\theta}$.

В качестве оценки неизвестного параметра $\hat{\theta}$ берется то его значение, при котором точное (теоретическое) значение первого момента совпадает с его империческим значением, найденным по

заданной выборке.

$$\begin{cases} M_1 = M_1(\theta) = \int\limits_{-\infty}^{+\infty} x f(x,\theta) dx, \\ \bar{x} = \frac{1}{n} \sum x_i, \\ M_2 = \bar{x}_2, \\ \vdots \\ M_k = \bar{x}_k \end{cases}$$

В случае многомерного параметра θ добавляются моменты более высоких порядков. Различают начальные и центральные моменты r-того порядка. Применение того или иного момента субъективно определяет исследователь. Чаще всего

$$\begin{cases} M_1(\alpha, \beta) = \bar{x}, \\ D(\alpha, \beta) = \hat{S}^2, \end{cases}$$

где $\hat{x}$ — теоретическое выборочное среднее, $D(\alpha, \beta)$ — теоретическая дисперсия, $M_1(\alpha, \beta)$ — мат. ожидание, $\hat{S}^2$ — эмпирическая дисперсия.

Метод моментов дает состоятельную оценку, но недостаточно эффективную. Альтернативой является *метод максимального правдоподобия.*

Пусть $x_1, x_2, \ldots, x_n$ — независимая случайная повторная выборка из некоторой генеральной совокупности с плотностью распределения $f(x, \theta)$ и $F(x, \theta)$. Пусть θ — единственный параметр. В этом случе строят *совместную* плотность распределения

$$f(x_1, \theta) \cdot f(x_2, \theta) \cdot \ldots \cdot f(x_n, \theta) = L(x_1, x_2, \ldots, x_n, \theta).$$

В качестве оценки θ берется значение параметра при котором функция, зависящая от элементов выборки и самого параметра, принимает максимальное значение. Уравнение максимального правдоподобия:

$$\frac{\partial L(x_1, x_2, \ldots, x_n, \theta)}{\partial \theta} = 0.$$

Для многих распределений задача имеет единственное решение, в других случаях метод неприменим. Распространяется на случий многих параметрических распределений. Из-за специфики функции максимального правдоподобия, как правило, рассматривают ее логарифм: $l \ln L$. Когда существует решение задачи, то полученная эффективная оценка является более эффективной по сравнению с методом моментов.

В случае многопараметрических распределений, условие превращается в систему (равенство нулю градиента).

Приведенные методы, таким образом, позволяют рассчитать (оценить) параметры распределений исследуемых случайных величин. В ряде случаев необходимо найти функцию связи между

двумя случайными величинами и оценить параметры этой функции, для чего используют модификации описанных методов. Кроме того, величины могут оказываться ненаблюдаемыми одновременно, что делает невозможным применение метода наименьших квадратов.

Методика контроля адекватности полученной функциональной зависимости:

1. необходимо провести проверку адекватности зависимости экспериментальных данных: если данные участвовали в установлении зависимости, то должна быть специально получена контрольная выборка среднего или большого размера;

2. строится две выборочные функции распределения, первая — для расчетной реализации ξ_1^* и контрольной реализации ξ_1^k;

3. полученные выборки сравниваются по известному критерию (в том числе по критерию Колмогорова Смирнова);

4. если контрольный параметр не превышает заданного порогового значения, то полученную функциональную зависимость можно использовать для прогноза величины ξ_1 по величине xi_2, в противном случае — нельзя.

11 Доверительные интервалы

Пусть $x_1, x_2, \ldots, x_n$ — независимая случайная повторная выборка из некоторой генеральной совокупности с плотностью распределения $f(x, \theta)$ и $F(x, \theta)$, зависящей от некоторого параметра θ.

Таким образом, необходимо оценить нижнюю $\underline{\theta}$ и верхнюю границы интервала $\bar{\theta}$, которому θ принадлежит с заданной вероятностью. $\underline{\theta}$ задается как $\underline{\theta}(x_1, x_2, \ldots, x_n)$, а $\bar{\theta}$ задается как $\bar{\theta}(x_1, x_2, \ldots, x_n)$.

То есть $\underline{\theta}$ и $\bar{\theta}$ — статистика. Интервал $[\underline{\theta}, \bar{\theta}]$ называется *доверительным интервалом* для параметра θ с коэффициентом доверия γ.

11.1 Методы построения доверительных интервалов

Одним из часто используемых методов является метод, основанный на *центральной статистике*. Центральная статистика — любая функция, зависящая от элементов выборки и параметра θ. Как правило выбирается функция монотонно убывающая или монотонно возрастающая. $K(\alpha)$ — *квантиль* уровня α. Рассмотрим функцию $T(x_1, x_2, \ldots, x_n, \theta)$ и

$$F(K_\alpha) = P\{T \le K\} = \alpha.$$

$$t_1 = K(\varepsilon_1), \quad t_2 = K(1 - \varepsilon_2).$$

Тогда

$$\begin{cases} P\{T \le t_1\} = F(t_1) = \varepsilon_1, \\ P\{T > t_2\} = 1 - F(t_2) = 1 - (1 - \varepsilon_2) = \varepsilon_2, \end{cases}$$

и

$$P\{t_1 < T(x_1, x_2, \ldots, x_n, \theta) \le t_2\} = 1 - \varepsilon_1 - \varepsilon_2 = r.$$

Если $\varepsilon_1 = \varepsilon_2 = \varepsilon$, то $r = 1 - 2\varepsilon$.

Лекция 2.04

12 Простая регрессия. Простой реляционный анализ

Пары точек с координатами (x, y) наносятся на координатную сетку. Необходимо получить предварительное представление о рассеивании точек. На графике возможны случаи:

1. нет ошибок в измерениях;
2. ошибки измерения незначительные;
3. присутствуют ошибки измерения и изменение результата, в связи с чем отсутствует явно выраженный тренд.

В ряде случаев возможно появление облака точек, где одному значению x соответствует несколько значений y и наоборот. В первой случае имеет функциональную зависимость, в остальных — стохастическую. Под функциональной зависимостью понимают соответствие некоторой величины $y \in Y$ единственной величине $x \in X$. Стохастическая зависимость предполагает что одному значению y, У может соответствовать несколько значений x_i, и наоборот. Во втором случае значения y рассматриваются как следствия x. То есть можно предположить, что есть некоторая функция, которая причинна для y по значению x. Таким образом необходимо оценить погрешность представления данных с целью выявления тренда (причинной функции).

Традиционно для оценивания корреляции двух случайных выборок используют коэффициент корреляции (Пирсена), который демонстврирует степень тесноты линейной зависимости. Судить о тесноте связи в случае нелинейного тренда оказывается ошибочным. Существуют различные характеристики связи и взаимосвязи выборок. Наиболее часто используется коэффициент корреляции, ковариации, корреляционная функция и взаимно корреляционая функция, нормированная корреляционная функция и взаимно нормированная корреляционная функция, моментная функция и взаимная моментная функция, функция регрессии, функция когерентности, условная плотность распределения вероятности, условная функция распределения вероятности, функция коллегации, коэффициент коллегации, коэффициент Спирмена и коэффициент Кэмбелла.

Коэффициент корреляции (Пирсена) определяется как отношение

$$r_{XY} = \frac{\mathbf{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}.$$

Таким образом, задача разбивается на две подзадачи:

1. установить вид тренда (линейный, нелинейный);
2. подтвердить тесноту связи.

Ковариация и коэффициент корреляции Пирсена устанавливают линейность функции связи (в противном случае считается что связь нелинейная).

Коэффициенты принимают значения близкие к максимальному только в случае тесной взаимосвязи. Установить тесную связь в случае нелинейной зависимости не представляется возможным. Если степень тесноты связи высокая, то можно говорить о построении *регрессии*.

Регрессия — условное математическое ожидание случайной переменной y при условии, что другая переменная приняла значение x. Если коэффициент корреляции Пирсена близок к единице, говорят о линейной регрессии. Моделью линейной регрессии является формализация, в которой мат. ожидание наблюдаемой величины y является линейной комбинацией независимых переменных. В случае $k > 1$ независимых переменных, говорят о *множественной* регрессии, при $k = 1$ о *линейной* регрессии.

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_kx_k.$$

Коэффициент a_0 называется *свободным* или *постоянных* членом, a_i — коэффициентами линейной регрессии.

Построение линейной регрессии лежит в основе корреляционного анализа. Корреляционный анализ изучает стохастическую зависимость между случайными величинами на основе совокупности коэффициентов и функций связи и взаимосвязи величин. При установлении нелинейной зависимости необходимо использовать альтернативные подходы, в том числе уже изученный подход на основе методов момента и максимального правдоподобия, а также на основе коэффициентов Спирмена и Кэнделла. Такие коэффициенты являются ранговыми.

13 Основные понятия теории проверки статгипотез.

Статистической гипотезой H называют любое утверждение, относительно функции распределения $F(x)$ некоторой случайной величины x, касающейся типа функции распределения, значения ее параметров и тд. Гипотезу H проверяют путем сопоставления выдвинутых предположений с результатами эксперимента, которые в статистике представляют собой n независимых наблюдений над случайной величиной x.

Гипотезу называют *простой*, если ее условиям удовлетворяет единственная функция распределения $F(x)$, в противном случае — *сложной*. Гипотеза, справедливость которой проверяется в ходе эксперимента, называется *основной* или *нулевой* гипотезой и обознается H_0. В зависимости от того, какие отклонения от H_0 возможны, формулируют *альтернативные* или *конкурирующие* гипотезы. *Статистические* гипотезы проверяют с помощью *статистического критерия*.

Статистический критерий — совокупность правил, позволяющих по полученной выборке принять H_0 и отвергнуть H_1 (она же H_a), либо наоборот — принять H_1 и отвергнуть H_0.

Опишем общий принцип построения критерия.

Задается некоторая функция $S(x_1, x_2, \ldots, x_n)$. Тогда множество всех возможностей значений S разбивается на два подмножества: T_0 и $T_{\text{кр.}}$ T_0 — основное множество или множество принятия решений H_0. $T_{\text{кр.}}$ — критическое множество, множество отвержения решений H_0.

Если конкретное значение статистики попадает в T_0, то гипотезу H_0 принимают, а H_1 отвергают. И наоборот.

Ошибка первого рода — если верна H_0, но принимается H_1.

Ошибка второго рода — если верна H_1, но принимается H_0.

Вероятность α ошибки первого рода называется *размером критерия* (размером критического множества). Вероятность β ошибки второго рода называется *значимостью критерия*. Часто используют также понятие *мощности критерия* $\gamma = 1 - \beta$.

Невозможно изменить обе ошибки одновременно. Как правило, фиксируют β и минимизируют α. Если критерий таков, что критическими являются как малые значения его статистики, так и большие, то критерий — *двусторонний*, в противном случае — *односторонний*. Выбор критерия определяется видом статистики.

Если множество основной и альтернативной гипотезы отделены третьим множеством (не все альтернативы учтены в постановке задачи), такое множество будем называть *зоной индифферентности*. Если заранее известен закон распределения случайной величины, то статистические выводы могут быть точнее.

При этом исследователь может ошибаться в выборе закона распределения. *Критериями согласия* называют критерии, в которых гипотеза определяет закон распределения полностью, либо с точностью до небольшого числа параметров. Пусть имеется выборка размера n с теоретической функцией распределения $F(x)$. Предположим, что найдена некоторая гипотетическая функция

распределения $G(x)$. Тогда гипотеза основная будет формулироваться как

$$H_0 : G(\bullet) = F(\bullet).$$

Если у нас достаточно большие n, стремящиеся к бесконечности, то $G_n(x) \to F_n(x)$.

Статистика Колмогорова:

$$D_n = \sup_x |F_n(x) - F(x)|.$$

<не разорбался, цитата>

Статистика Колмогорова: можно сравнивать Fn(x) и F(x) в равномерной метрике (см фото). Статистику называют статистикой Колмогорова, если гипотеза верна, то Fn(x) -> F(x), в противном случае отвергается в пользу альтернативного распределения. Величина табулирована в процентах таблицах до значения n = 35, выбирается уровень значимости критерия объем выборки, на пересечении в таблице - пороговое значение. В качестве альтернативы для n >= 35 есть формула (в самом низу доски). Альтернативой критерию является критерий мат статистики Омега2. Состоятельность критерия Колмогорова и Омера:2 означает что любое отличие распределения выборки от теоретического будет с их помощью обнаружено, если наблюдения будут продолжаться достаточно долго. Более трудоемкой оказывается проверка сложной гипотезы. Например, требуется подтвердить наличие распределения. Необходимо на основе метода максимального правдоподобия получить некоторую оценку неизвестных параметров, после чего воспользоваться модифицированной формулой две звездочки(самый низ справа на доске)

</не разорбался, цитата>

В ряде случаев удобно работать не с конкретными значениями признака объекта, а с группами значений одного ранга. Ранг как правило выбирается исследователем (как и его интерпретация), в силу чего недостатком подхода является субъективность входных данных, при этом подход упрощает обработку данных и демонстрирует более объективные результаты анализа. Примером ранговой корреляции является ранговая корреляция Спирмена и Пирсена.

13.1 Метод Спирмена

Пример. Даны оценки по защите дипломных проектов в группе n вуза Y, данные по оценке абитуриентов (средний балл, полученный студентом группы n на вступительных экзаменах).

Ставится задача проверить статистическую гипотезу. Основная задача — признаки независимы. Альтернативная гипотеза — признаки зависимы.

Спирмен предложил статистику следующего вида:

$$\rho_s = 1 - \frac{6}{n(n-1)(n+1)} \sum_{i=1}^{n} (R_i - S_i)^2.$$

Статистика $\sqrt{n-1}\cdot\rho_s$ распределена по нормальному закону. Так как величина больше 1.64, мы оказываемся в критической области, т.е. отвергаем нулевую гипотезу при ошибке второго рода.

Ранговый коэффициент Спирмена прост в реализации, при этом позволяет устанавливать тесноту связи только для пары переменных.

Статистика χ^2, предложенная Пирсоном, вычисляется по формуле

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(O_i/N - p_i)^2}{p_i}.$$

Строится график распределения χ^2, границы критической области табулированы и находятся в таблицах ранговой корреляции Пирсена на пересечении строк уровня значимости и степени свободы распределения χ^2. Статистика χ^2 имеет степень свободы $(k-1)(m-1)$, где k — количество строк, m — количество столбцов. Таким образом, подтверждается гипотеза о независимости признаков.

14 Методы многомерного статистического анализа

Каждый объект выборки может содержать наблюдение более чем над одной случайной переменной (задача множественной регрессии, когда все переменные считаются случайными). В этом случае изучается взаимосвязь между зависимой переменной с одной стороны и набором независимых переменных с другой стороны. В других методах многомерного анализа все случайные переменные анализируются одновременно как случайный вектор, имеющий многомерное распределение.

Многомерный статистический анализ — анализ множественных результатов измерений свойств случайной выборки, в рамках которого используются множество методов решения целого спектра задач.

14.1 Анализ временных рядов

Временной ряд — совокупность последовательных значений переменных (процессы), полученных в результате эксперимента, через определенные, чаще равные, значения базового параметра. АВР используется для решения следующего класса задач:

1. как математическая модель процессов представлена временным рядом;
2. с целью исследования структуры временного ряда для выявления тренда процесса и обнаружения периодических колебаний;
3. с целью прогнозирования будущего развития процесса, представленного временным рядом;
4. для исследования взаимодействия между различными временными рядами.

Для решения этих и подобных задач используют:

1. методы корреляционного анализа (выявляются наиболее существенные периодические и квазипериодические зависимости и их задержки (лаги) в одном процессе или между несколькими процессами);
2. методы сглаживания, фильтрации для преобразования временных рядов с целью удаления выбросов;
3. методы авторегрессии и скользящего среднего для описания и прогнозирования процессов, проявляющих однородные колебания вокруг некоторого среднего значения.

При исследовании сложных объектов и систем часто нет возможности измерить исследуемую величину, в связи с чем возникает необходимость расчета этой величины по значениям двух и более измеряемых величин, определяющих свойства объекта или системы исследования (факторы).

При этом для измерения доступной величины в той или иной степени влияющей на результат (степень влияния неизвестна), кроме того влияние неизвестного фактора проявляется в нескольких признаках, то есть признаки могут обнаруживать тесную связь между собой (коррелируемость).

`ну алло`

Исследователь может существенно сократить число факторов, влияющих на решение задачи, тем самым снизив размерность задачи. Для обнаружения влияющих на измеряемые переменные факторов используются методы *факторного анализа*.

Факторный анализ. Часто используется выбор новых признаков, являющихся линейными комбинациями прежних и «вбирающих в себя» большую часть изменчивости наблюдаемых данных, и поэтому передающих большую часть информации, заключенную в первоначальных наблюдениях. Обычно для этого используют *метод главных компонент*.

Этот метод сводится к выбору новой ортогональной системы координат в пространстве наблюдений. В качестве первой главной компоненты выбирают направление, вдоль которого массив наблюдений имеет наибольший разброс. Последующая главная компонента выбирается таким образом, чтобы разброс наблюдений вдоль нее оказался максимальным и она была ортогональна другим компонентам, выбранным вдоль нее.

Недостатком подхода является переход от интерпретируемых факторов к факторам, которые сложно или невозможно интерпретировать. Суммарный вклад отброшенных факторов в решении задачи может превосходить два главных компонента (второй недостаток).

Причинно-следственный анализ. Также как факторный анализ помогает уменьшить количество факторов, при этом установив причинно-следственную связь между парами факторов. Отличается от факторного анализа последовательностью связей функциональных зависимостей, часто используется в психологии.

Дискриминантный анализ. Если необходимо построить функцию измеряемых характеристик, значение которой позволит разбить исходные данные на две группы, используют *дискриминантный анализ*. Как правило говорят о *линейном* дискриминантном анализе, в котором классифицирующие признаки выбираются как линейные функции от первичных признаков. Результаты дискриминантного анализа легко интерпретируются, так как дискриминирующая функция строится по экспериментальным данным как множественная регрессия. Коэффициенты регрессии иллюстрируют вклад каждой независимой переменной в решении задачи. Если вновь появившийся объект имеет значение функции y^*, превышающее величину регрессии, то объект может быть однозначно помещен в *базовую* группу. В противном случае — в *альтернативную* группу. Базовой является группа объектов с соответствующими точками выше плоскости регрессии.

Кластерный анализ. Если число групп больше двух, то применяются методы *кластерного анализа.* Кластеризация предполагает объедение объектов со схожими свойствами в одной группе (кластере), смысл которой хорошо интерпретируется. Большинство методов кластеризации относятся к иерархическим, то есть построенных в соответствии с дендрограммой (деревом). Иногда реализуется обратная схема, когда все объекты размещены в один кластер нулевого уровня, который на основе выбранных условий разбивается на два, а при необходимости больше кластеров (дивизивная схема). Выделяют другой класс методов неирархической кластеризации, когда оценивается близость объектов к выбранному или заданному центру кластера. Общим недостатком всех методов кластеризации является субъективность решения.

Часто в качестве исходных данных используются не сами оценки степени сходства объектов, а результаты их ранжирования. Если попарно сравнивая объекты из выборки мы можем установить более высокий ранг популярным объектам, а самый низкий непопулярным, то говорят о решении задачи методом шкалирования (н: оценки политических лидеров).

15 Моделирование случайных величин

Первые алгоритмы генерации случайных чисел были предложены американскими учеными Нейманом и Уламом с целью моделирования поведения нейтронов в атомном реакторе. Выделяют:

1. физические ГСЧ.
2. алгоритмические ГСЧ;
3. алгебраические ГСЧ.

Физические ГСЧ. Используют случайную природу регистрируемых случайных величин (орел-решка, таймер, вольтметр, амперметр). Достоинство — высокая точность генерируемых значений. Недостаток — необходимость подтвержения распределения выбранной физической величины в соответствии с заданных законом распределения выбранной случайной величины, необходимость обеспеченивания интерфейса между прибором, регистрирующим случайную величину и компьютером.

Алгебраические ГСЧ. Альтернативной физическим являются алгебраические ГСЧ. Предполагается что случайные величины в природе распределены по законам, алгебраически связанными с равномерным распределенем случайных величин в интервале $[0, 1]$. В этом случае равномерное распределение в интервале $[0, 1]$ является базовым, и на его основе получаются СВ с заданными законами распределения.

Пример. Джон фон Нейман привёл следующий метод получения десятизначных псевдослучайных чисел: «Десятизначное число возводится в квадрат, затем из середины квадрата числа берётся десятизначное число, которое снова возводится в квадрат, и так далее». Например (для простоты возьмем четырехзначные числа):

$$3471 \to 47^2 = 2209 \to 20^2 = 4000.$$

Такой датчик имел небольшой период зацикливания либо приводил к генерации нулевых значений.

В 1949 году американский математик Дерик Клемер предложил *линейный конгруэнтный метод* для генерации случайных чисел. Каждую новую случайную величину находили как

$$x_{n+1} = (ax_n + c) \mod m,$$

где $m \geq 2,\ 0 \leq a \leq m,\ 0 \leq c \leq m,\ 0 \leq x_0 \leq m$.

</КОНСПЕКТЫ НАТАШИ>

16 Кластерный анализ

Разделяется на иерархические и неиерархические методы. В иерархических методах строится дедраграмма; объединяются кластеры; деление.

Пусть $\{x_1, \ldots, x_n\}$ — множество объектов $\{y_1, \ldots, y_m\}$ — множество меток (названий кластеров). Каждый объект выборки $x \in X$ может характеризоваться набором координат. Необходимо разбить исходную выборку x на **непересекающиеся** подмножества, называемые *кластерами*. Кластер состоит из объектов, близких по некоторой метрике d, выбранной исследователем. При этом объекты разных кластеров существенно отличаются.

Предположим, что имеется конечная обучающая выборка объектов, на основе которой устанавливаются условия вхождения вновь прибывшего объекта в кластер. Алгоритм кластеризации — это функция поиска соответствия элементов множеств x и y. Количество кластеров может быть предварительно задано либо изменяться в процессе кластеризации. Как правило, если количество кластеров задано и они интепретируемы, то говорят о задаче *классификации (обучения с учителем)*. В противном случае говорят об обучении без учителя или кластеризации.

Кластеризация дает принципиально неоднозначное решение, так как:

1. не существует однозначно наилучшего критерия качества кластеризации (сформулирован ряд эвристических критериев, основанный на результатах классификации аналогов);
2. существуют различия в стандартизации переменных;
3. число кластеров устанавливается исследователем субъективно;
4. результат кластеризации зависит от метрики.

Кластер имеет следующие математические характеристики:

1. центр — среднее арифмитическое место точек в пространстве переменных;
2. радиус — максимальное расстояние от точки до центра кластера (с учетом выбранной метрики);
3. среднее квадратическое отклонения от центра;
4. размер кластера.

СКО вычисляется на основе евклидова расстояния между точкой и центром и определяется по формуле

$$\hat{\sigma} = \frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{N} (x_{ij} - \bar{x}_{ij})^2$$

Размер кластера определяется либо по радиусу кластера, либо величиной, превыщающей СКО.

По результатам кластеризации можно получить две группы с перекрытием. Объекты, являющиеся пересечением двух кластеров называются *спорными*, но должны быть отнесены к одному из кластеров однозначно.

Критерием для определения схожести и различия кластеров является сравнение с количественной мерой, которая, как правило, вычисляется в евклидовой метрике.

16.1 Иерархические и неиерархические методы кластеризации

Так как расстояние между объектами отражает меру сходства, то выбранная метрика должна удовлетворять следующему условию:

1. $d_{ij} \geq 0$,
2. $d_{ij} \equiv d_{ji}$,
3. $d_{ij} \leq d_{ik} + d_{kj}$,
4. $d_{ij} \implies 0 \iff i \neq j$,
5. иногда вводят условие $d_{ij} = 0 \implies i = j$.

16.1.1 Некоторые примеры метрик

Метрика Минковского:

$$d_{ij} = \left(\sum_{k=1}^{n} |x_{ik} - x_{jk}|^{\lambda} \right)^{\frac{1}{\lambda}}.$$

Метрика Канберра:

$$d_{ij} = \sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}.$$

Метрика (расстояние) Варда:

$$d_{ij} = \sum_{i=1}^{N_i} \left(x_i - \frac{1}{N_j} \sum_{j=1}^{N_j} x_j \right).$$

16.2 Оценка качества кластеризации

1. Ручная проверка;
2. установка контрольных точек и проверка на полученных данных;
3. определение стабильности кластеризации путем добавления в модель новых данных;

4. создание и сравнение кластеров с использованием разных способов кластеризации

На практике для больших объемов данных используется последний подход. При малом объеме данных же используются первые два.

16.3 Метод к-средних

В дз делаем для 2, 3, 4 кластеров ($m = 2, 3, 4$). Алгоритм:

1. считается расстояние каждой точки выборки до центра предполагаемого кластера;
2. то же самое и выбираются точки, близкие к первому, второму, третьему центру, соответстенно;
3. точки распределяются по кластерам;
4. вновь рассчитывается центры кластеров (по определению центра).

Процедура повторяется до тех пор, пока точки не стабилизируются в кластерах на протяжении трех последних итераций.

Достоинства метода к-средних —

1. простота алгоритма;
2. высокое быстродействие;
3. не тербует допольнительных сведений о системе.

Недостатки метода к-средних:

1. может медленно работать на больших объемах данных;
2. чувствителен к выбросам;
3. в качестве начальных центроидов выбираются произвольные точки, что влияет на время вычилсения и точность результата.

16.4 Метод Варда

Метод иерархический агломеративный. Начальными центроидами являются все объекты исследования. В качестве расстояния между кластерами берется прирост суммы квадратов расстояний от объектов до центров кластеров, получаемых в процессе объединения.

$$d_x = \sum_{j=1}^{N_Y} \left(x_i - \frac{1}{N_X} \sum_{i=1}^{N_X} x_i \right)^2 .$$

Вычисляются промежуточные величины d_x, d_y и, в предположении что XY является объединением двух кластеров предыдущего этапа X и Y. Тогда метрикой метода является величина

$$d = d_{XY} - (d_x + d_y).$$

Метод направлен на объединение близко расположенных кластеров и позволяет однозначно сохранять далеко расположенные кластеры.

Достоинства метода Варда:

1. однозначно определяет начальные центроиды;
2. метрика метода задана однозначно $\Longrightarrow$ дает однозначную кластеризацию.

Недостатки метода Варда:

1. невысокое быстродействие;
2. по прежнему чувствителен к выборсам.

16.5 Метод ближайшего соседа

Также известен «одиночная связь». Является иерархическим агломеративным. Определяется расстоянием между двумя наиболее близкими объектами в соседних кластерах.

Метод позволяет выделать классы сколь угодно сложной формы, представляемые в виде длинных цепочек («волокнистый кластер»).

16.6 Метод удаленного соседа

Встречается под названием «полная связь». Относится к иерархическим агломеративным методам. Расстояние между двумя кластерами определяется как максимальное между двумя объектами кластеров. Метод хорошо использовать для классификации объектов из разных «рощ». Если кластеры имеют цепочечную форму — метод использовать не стоит.

Дивизивные методы являются обратными к конгламеративным и используют тот же самый подход, но уже с целью разделения кластера на два и больше. Уступают в быстродействии агломеративным, но применимы если разделение производится на незначительное количество групп (кластеров).

17 Дисперсионный анализ

Метод наименьших квадратов плюс систематизированный набор правил для проверки статистических гипотез носит обобщенное название *дисперсионный анализ*. Анализируется величина

$$y = \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

Эта формула разбивается на суммы квадратов, связанных с результатами действия различных классификационных факторов, определяемых индексами данных.

$$\sum_{ij} (y_{ij} - y_{\bullet\bullet})^2 =$$

$$= \sum_{ij} (y_{i\bullet} - y_{\bullet\bullet})^2 + \sum_{ij} (y_{\bullet j} - y_{\bullet\bullet})^2 + \sum_{ij} (y_{i\bullet} + y_{\bullet j} - y_{\bullet})^2 ,$$

где первое слагаемое соответствует основному эффекту параметра A, второе — основному эффекту параметра B, а третье - эффекту взаимодействия параметров, $\cdot$ — результат осреднения величины y_{ij} по соответствующим индексам.

Путем анализа данных с помощью приведенной модели можно выделить наиболее сильный фактор. Если максимальной получилась величина от эффекта взаимодействия A и B, проверяется гипотеза с другим набором факторов.

Метод оказывается информативным в плане выделения значимых факторов, при этом требует проведения (часто масштабного) натурного эксперимента.

В некоторых случаях данные поступают последовательно, в связи с чем проверка гипотезы о влиянии того или иного фактора также производится последовательно.

Пусть имеются выборки двух показателей: $(x_1, x_2, \ldots, x_n)$ и $(y_1, y_2, \ldots, y_n)$ с неизвестными параметрами $N(\mu_1, \sigma_1)$ и $N(\mu_2, \sigma_2)$ нормальных распределений. Необходимо подтвердить или опровергнуть гипотезу о равенстве матожиданий исследуемых выборок. В этом случае можно обобщить результаты диспресионного анализа на вновь полученную выборку. Вычисляются выборочные средние

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{m} y_i$$

и выборочные квадраты

$$S_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

$$S_y^2 = \frac{1}{m} \sum_{i=1}^{m} (y_i - \bar{y})^2.$$

Далее получают результирующую оценку общей дисперсии

$$S^2 = S_x^2 + S_y^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{m}\sum_{i=1}^{m}(y_i - \bar{y})^2$$

$$mn \cdot S^2 = m\sum_{i=1}^{n}(x_i - \bar{x})^2 + n\sum_{i=1}^{m}(y_i - \bar{y})^2.$$

Необходимо оценить выборочную дисперсию S_{x-y}^2 (используется распределение и статистика Стьюдента). Также нужно ввести статистику

$$t = \frac{\bar{x} - \bar{y}}{S}\sqrt{\frac{mn}{m+n}}.$$

Статистика подчиняется распределению Стьюдента с $m + n - 2$ степенями свободы.

Из статистических таблиц, в соответствии со значением степени свободы и выбранного уровня значимости полученного решения, получается оценка границ интервала $[t_-, t^-]$. Также оценивается величина $\hat{t}$ и принадлежность этой величины к найденному отрезку. Если принадлежит, то говорят о совпадении матожидания при совпадении СКО σ, что позволит перенести выводы, полученные по выборке X на выборку Y.

Результат сравнения двух выборок сводят в так называемую *таблицу диспресионного анализа*:

Источник изменений	*Суммы квадратов*	*Степени свободы*
различия между выборками	$\frac{mn}{m+n}(\bar{x}-\bar{y})^2$	1
различия внутри выборок	$\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{j=1}^{m}(y_j-\bar{y})^2$	m + n - 2
полная изменчивость	$\sum_{k=1}^{L}(z_k-\bar{z})^2$	L

Выборку Z считают объединением по всем выборкам, к примеру:

$$X=(x_1,x_2,x_3),\quad Y=(y_1,y_2),\quad Q=(q_1,q_2)$$

$$Z=(x_1,x_2,x_3,y_1,y_2,q_1,q_2).$$

Анализируя ТДА, в случае значительных различий между выборками и незначительного внутри выборок, а также незначительном влиянии межфакторного взаимодейтсвия, можно принять гипотезу о разделении величин на два независимых кластера X и Y.

<u>*Важно*</u>: то есть, дисперсионный анализ может быть применим в двух случаях, для анализа главных факторов, и для разделения исходных объектов на группы.

Если требуется сравнить три и более обработок, то формулы и таблицы меняются следующим образом:

Источник изменений	*Суммы квадратов*
различия между выборками	$b=\sum_k n_k x_{\bullet k}^2-nx_{\bullet\bullet}^2$
различия внутри выборок	$a=\sum_a\sum_r(x_{rs}-x_{\bullet\bullet})^2=\sum_s\sum_r(x_{rs}-x_{\bullet s})^2+$ $+\sum_s n_s(x_{\bullet s}-x_{\bullet\bullet})^2$
полная изменчивость	$\sum_{i=1}^{N}(z_i-\bar{z})^2$, где $N=\sum_{p=1}^{K}n_p$

где k — номер выборки, точки определяют осреднения по соответствующим коэффициентам.

Достоинством дисперсионного анализа является универсальность по количеству сравниваемых выборок, кроме того, дисперсионный анализ позволяет как выявить главные факторы, так и разделить исследуемые объекты на группы. Недостаток — сложен в понимании и требует внимательности от исследователя в смысле интерпретации главных факторов и формирования сравниваемых выборок.

На практике вычисления проводят по следующему методу:

$$\begin{cases} c_1 = \sum_{i=1}^{n_1} n_{1i} x_i, \\ c_2 = \sum_{i=1}^{n_2} n_{2i} x_i, \\ \vdots \\ c_k = \sum_{i=1}^{n_k} n_{ki} x_i \end{cases} \quad \Longrightarrow \; C = \sum_{j=1} \sum_{i=1} n_{ij} x_i.$$

$\bar{x}_i = c_i/n_i$, и общее среднее $\bar{g} = S/N$, $N = \sum n_i$. Тогда анализируется величина, (условно) соответствующая межвыборочной сумме квадратов:

$$\sum_{i=1}^{k} c_i \bar{x}_i - S\bar{g},$$

и/или внутривыборочная сумма квадратов

$$\sum_{j=1} \sum_{i=1}^{k} c_i (x_j i - \bar{x}_i)^2 - \bar{g}^2.$$

18 Факторый анализ

Факторный анализ — многомерный метод, применяемый для изучения взаимосвязей между значениями исследуемых переменных. Позволяет снизить размерность задачи за счет сокращения факторов исследования.

Анализируя результаты лабораторной работы 4 (метро), можно сделать вывод, что результат решения задачи оказывается очень чувтсвительным к масштабу входных данных. Кроме того, при получении функциональных зависимостей между исследуемыми параметрами, замена слабонелинейной функции связи на линейную также приводила к значительному изменению результата исследования.

Для коррекции модели необходимо использовать методы кластерного анализа. Факторый анализ позволяет решить две важные проблемы:

1. выявить скрытые факторы, воздействие которых определяет наличие нелинейности с корреляцией между исследуемыми переменными;

2. исключить из исследования факторы, влияющие на решение задачи и делающие её чувствительной к изменению входных данных.

Таким образом, задачи с неустойчивым решением к концу 20 в. стали решаться методами, ориентированными либо на уделение фактора, обеспечивающего неустойчивость, либо эксперимента или конкретного испытания в эксперименте, где аномальное значение этого фактора приводит к неустойчивости решения.

Инструментами факторного анализа, таким образом, являются: *метод главных компонент, дисперсионный анализ, корреляционный анализ.*

Если существенным является взаимодействие двух факторов, а сами выделенные факторы мало влияют на изменение решения, то, скорее всего, существуют иные факторы, которые не были учтены. Эта процедура позволяет выявить латентные (скрытые) переменные, что увеличивает размерность задачи, но в целом повышает устойчивость решения:

$$y = f(x)$$

$$y = a_0 + a_1 x_1 \longrightarrow y = a_0 + a_1 x_1 + a_2 x_2.$$

Тем не менее, после ряда включений-исключений факторов, окончательное их количество определяет *метод главных компонент.* Это **единственный** математически обоснованный выделения значимых факторов.

Следовательно, факторный метод может быть разведочным (исследует латентную структуру исследуемого объекта или системы) или конфирматорным (подтверждающим — предназначенным для подтверждения гипотез о факторах и их конкретных значениях).

19 Анализ выборосов

Различают понятия *цензурирования* исходных выборок и, как таковое, выявление *выборосов* — аномальных значений, значимо влияющих на решение задач.

Существует несколько подходов.

19.1 Предварительный анализ

Одно из значений значимо отличается от других элементов выборки или существенно меняет величину выборочного среднего или выборочной дисперсии. Чем равномренее распределение значений на интервале изменения случайной величины x тем симметричнее функция распределение $f(x)$ (?что?). Поэтому рассмтаривают расстояния между медианой и квантилем уровня 0.25 (первым квантилем), и медианой и квантилем уровня 0.75 (третьим квантилем). Если расстояния существенно отличаются, то можно говорить о наличии выборосов. Другой характеристикой является расстояние между первым и третьим квантилем. Если ожидается получение значений в заданном диапазоне, то можно априорно оценить величину, которую не привысит такое межквантильное расстояние. Если эта величина превышена, то имеет место выброс.

19.2 Цензурирование

Возможно также цензурирование слева: например, только положительные значения.

Важно: таким образом, нельзя безосновательно исключать выбросы из выборки. Можно лишь изменить область определения исследуемой случайной величины.