О ЛИНЕЙНОЙ АППРОКСИМАЦИИ ВЕКТОРНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ

Муха В. С., Будный Р. И.
Кафедра информационных технологий автоматизированных систем, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: mukha@bsuir.by, budnyjj@gmail.com

Рассматривается проблема линейной аппроксимации векторных статистических данных. Обсуждаются области применения, объекты, цели и критерии аппроксимации. Основное внимание уделяется традиционной аппроксимации линейной регрессией с критерием минимума суммы квадратов вертикальных расстояний и симметричной аппроксимации с критерием минимума суммы квадратов перпендикулярных расстояний (по К. Пирсону). Приводятся результаты компьютерного статистического моделирования и теоретические аргументы, определяющие области предпочтительного применения этих видов аппроксимаций.

Введение

Под аппроксимацией статистических данных будем понимать их замену линейной детерминированной зависимостью, близкой к этим данным в каком-то смысле. Аппроксимация широко применяется в регрессионном анализе, теории планирования эксперимента, идентификации объектов и систем, распознавании графических представлений объектов и текста. В различных приложениях аппроксимация интерпретируется по-разному, чаще всего как математическая модель объекта (системы) или как преобразование геометрических фигур. В настоящее время в этой области имеется ряд вопросов, на которые нет окончательных ответов. В данном докладе предполагается заполнить некоторые существующие пробелы в этой области.

I. Классификация стохастических объектов

Векторные статистические данные могут порождаться объектами (системами) следующих трех видов.

Полустохастическим объектом будем называть объект с детерминированным входом и случайным выходом. Это либо регрессионный объект (объект с внутренним шумом на выходе), либо детерминированный объект с ошибками в измерениях выходных переменных (рисунок 1). Объект описывается условной плотностью вероятности $f(\eta/\xi,\theta)$, η – выходная переменная, ξ – входная переменная, θ – параметр объекта, y – наблюдение выходной переменной, $f(y/\eta)$ – условная плотность вероятности, описывающая измерительную систему, ВУ – вычислительное устройство, $\Delta(\xi,y)$ – результат аппроксимации. Входная и выходная переменные, а также параметр могут быть многомерными.

Стохастическим объектом 1-го типа будем называть объект со случайными входом и выходом (рисунок 2). Такой объект описывается совместной плотностью вероятности $f(\xi,\eta)$. Возможно наличие ошибок в измерениях входных и выходных переменных.

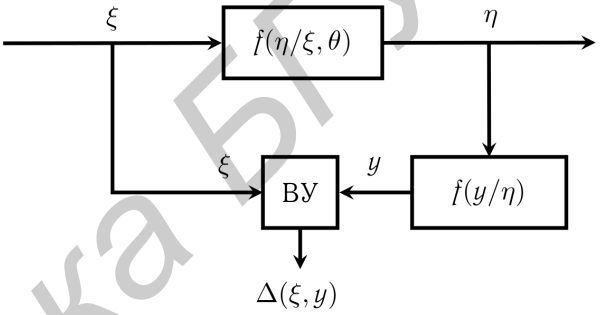

Рис. 1 – Схема полустохастического объекта

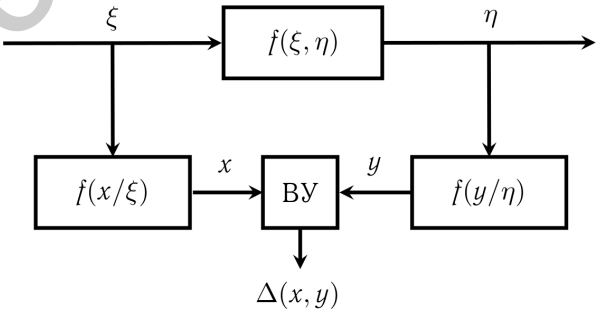

Рис. 2 – Схема стохастического объекта 1-го типа

Стохастическим объектом 2-го типа назовем детерминированный объект с ошибками в измерениях входных и выходных переменных (рисунок 3). Он описывается детерминированной зависимостью $\eta = \phi(\xi,\theta)$, θ – вектор параметров объекта.

II. Линейная аппроксимация стохастических объектов

Очевидно, целью аппроксимации любого из перечисленных выше объектов является получение возможности предсказания состояния (выходной переменной) объекта по наблюдению входной переменной, так что критерии оптимальности аппроксимации должны ориентироваться на эту цель. В настоящее время наиболее

широкое распространение получил критерий минимума суммы квадратов вертикальных расстояний от наблюдений выходной переменной до аппроксимирующей прямой или плоскости (классический критерий наименьших квадратов). Этот критерий формально применим ко всем видам рассмотренных выше объектов.

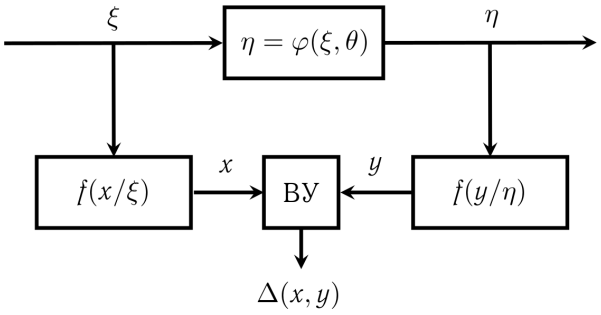

Рис. 3 – Схема стохастического объекта 2-го типа

Известен также иной критерий, состоящий в минимизации суммы квадратов перпендикулярных расстояний от наблюдений входных и выходных переменных до аппроксимирующей прямой или плоскости [1, 2]. Линейная аппроксимация с данным критерием в [2] названа симметричной.

Области предпочтительного использования названных выше критериев достаточно четко не определены. Этот вопрос является предметом исследований данной работы.

Относительно задачи линейной аппроксимации полустохастического объекта можно сказать, что она полностью решена в рамках классического линейного регрессионного анализа как задача оценивания параметров линейной математической модели объекта с критерием минимизации суммы квадратов вертикальных расстояний. Оптимальное решение здесь – классическая линейная регрессия, обеспечивающая как оптиммальное оценивание параметров математической модели объекта (при заданных значениях входных переменных), так и оптимальное прогнозирование наблюдений выходных переменных по наблюдениям входных переменных.

III. Компьютерное моделироование

К стохастическим объектам 1-го и 2-го типов формально можно применять любой из названных выше критериев оптимальности. Однако с точки зрения эффекта такого применения эти объекты следует различать.

Так как стохастический объект 1-го типа не имеет реальных параметров, то сравнение критериев по точности оценивания параметров линейной модели для такого объекта не имеет смысла. Смысл имеет сравнение по точности прогнозирования наблюдения выходных переменных по наблюдению входных переменных. Компьютерное моделирование такого объекта со скалярным входом и скалярным выходом и совместным нормальным распределением входа и выхода и ошибками в измерениях входных и выходных переменных показало, что более точное прогнозирование, как и для полустохастического объекта, обеспечивается классической линейной регрессией. Данный вывод соответствует теоретическим результатам работ [3, 4], в которых сформулирована задача оптимального прогнозирова-ния выходных переменных и показано, что оптимальным линейным предиктором является классическая линейная регрессия.

Моделировался также стохастический объект 2-го типа – детерминированный объект со скалярными входом и выходом, линейной зависимостью $\eta = \alpha + \beta\xi$ и ошибками в измерениях входной и выходной переменных. Нас здесь могут интересовать как точность оценок параметров объекта, так и точность предсказания наблюдения выхода по наблюдению входа. Моделирование показало, что точность оценивания параметров зависит от средних квадратических ошибок (с.к.о.) измерений входной и выходной переменных и от величины коэффициента усиления β. Для принятия решения о том, какой метод следует применять для получения более точных оценок параметров, предлагается следующее эмпирическое правило:

$$\sigma_\delta \le (0{,}7 + |\beta|)\sigma_\epsilon, \qquad (1)$$

где σ_ε – с.к.о. ошибок измерений входной переменной, σ_δ – с.к.о. ошибок измерений выходной переменной. Если условие (1) выполняется, то симметричная аппроксимация дает более точные оценки параметров, чем классическая линейная регрессия. В противном случае использование классической регрессии является более предпочтительным. Вместе с тем моделирование показало, что чем более точно метод оценивает параметры объекта, тем менее точно он оценивает (прогнозирует) состояние объекта.

IV. Список литературы

1. Pearson, K. On lines and planes of closest fit to systems of points in space / K. Pearson / Philosophical Magazine. – 1901. – V. VI. – N 2. – P. 559 – 572.
2. Муха, В. С. Симметричная аппроксимация векторных статистических данных линейными многообразиями / В. С. Муха // Весці Нац. акад. навук Беларусі. Сер. фіз.-мат. навук. – 2016. – № 4. – С. 23 – 31.
3. Муха, В. С. Оптимальные статистические решения для непрерывных многомерно-матричных состояний и наблюдений / В. С. Муха // Весці Нац. акад. навук Беларусі. Сер. фіз.-мат. навук. – 2010. – № 3. – С. 17 – 24.
4. Муха, В. С. Минимальный средний риск и эффективность оптимального полиномиального многомерно-матричного предиктора / В. С. Муха // Кибернетика и системный анализ. – № 2. – 2011. – С. 121 – 130.