Непараметрическая имитационная модель закона распространения ошибок

В. Л. Горохов[1], И. А. Брусакова[1], Р. И. Гайнутдинов[2]
[1]*Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)*
[2]*Санкт-Петербургское отделение Специальной астрофизической обсерватории РАН*

***Аннотация*. Рассматриваются методы и средства учета погрешностей косвенных многомерных измерений характеристик объектов мониторинга в условиях априорных неопределенностей в отношении распределений этих измерений.**

Ключевые слова: погрешности косвенных измерений; метод Монте-Карло; непараметрическая статистика

I. Постановка задачи. Введение

В традиционной теории погрешностей (ошибок) измеренная величина x может быть представлена как величина случайная и она подчиняется нормальному распределению с математическим ожиданием μ и среднеквадратическим отклонением σ (дисперсия σ^2) [1]. Эта величина σ называется погрешностью (ошибкой) измерений. Еще одна модель погрешностей в классической теории ошибок представленная в виде $\Delta x=x_{ист}-x_{изм}$ сводиться к выше означенной [1]. В практике физических экспериментов часто измеряются несколько разных физических величин $(x_1,..,x_n)$ и, у них вычисляется свои σ_{xn} , а затем выполняются операции с этими величинами, в результате получаем косвенные измерения: $\mathbf{X}_n=f(x_1,..,x_n)$.

Теперь, возникает вопрос о том какова погрешность этих косвенных измерений. Стандартная формула расчёта ошибки косвенного измерения выводится из простых соображений. Пусть у нас есть n нормально-распределённых случайных величин $\{x_i\}$ (i=1,..,n) с измеренными математическим ожиданиями μ_n и дисперсиями $\{\sigma^2_i\}$; (i=1,..,n) и ковариациями. Пусть также задана линейная функция от этих величин: $f(x_1,...x_n)=\mathbf{X_n}$. Как известно, значение $\mathbf{X_n}$ также будет иметь нормальное распределение. Чему будет равна погрешность такого косвенного измерения. Это хорошо известный результат теории ошибок и формула вычисления этой погрешности косвенного измерения известен как **закон распространения ошибок (**закон распространения погрешностей) [1].

Для случая нормального распределения прямых измерений и линейной функции формирования косвенных измерений $f(x_1,...,x_n)=\Sigma a_i x_i$ можно трактовать стандартное отклонение как погрешность косвенного измерения Δf:

$$\Delta f = \sqrt{[a^2_1 \Delta x^2_1 +..+a^2_n \Delta x^2_n + ... + 2a_1 a_2 \operatorname{cov}(x_2 x_1) +\ldots]} \quad (1)$$

В случае, если функция нелинейная, её представляют в виде разложения в ряд Тейлора до первого порядка:

$$\Delta f=\sqrt{[(\partial f/\partial x_1)^2\Delta x^2_1+,\ldots,+(\partial f/\partial x_n)^2\Delta x^2_n+2(\partial f/\partial x_1)(\partial f/\partial x_2)\operatorname{cov}(x_1x_2)+,\ldots]} \quad (2)$$

Это и есть стандартная формула (2) расчёта погрешности косвенных измерений – **закон распространения ошибок**.

К сожалению, такой подход обладает ограниченной применимостью в связи со следующими особенностями реальными эксперимента:

- Измеренные значения должны описываться нормальным распределением и только тогда ошибки трактуются как стандартные отклонения этих случайных величин. Если исходные измерения не подчиняются нормальному распределению, т. е. если прямые измерения не обладают нормальностью, то и измерения погрешности не могут быть представлены оценкой σ! Соответственно и для оценки погрешностей косвенных измерений стандартный закон распространения ошибок не работает!
- Кроме того, необходимо, чтобы ошибки были достаточно малыми, а функции достаточно гладкими, чтобы неточность из-за представления функции в линеаризованном виде была незначительной.
- В реальной практике экспериментальные прямые измерения могут иметь самые разные распределения (например: распределение Коши, Гамма распределение, распределение Релея–Райса и т. д.), Эти распределения могут быть несимметричными. Кроме того на практике погрешности прямых измерений по величине могут иметь разную величину (неоднородность).
- Особенно важно подчеркнуть, что в реальной практике эксперимента характер зависимости косвенных измерений от прямых измерений далёк от линейности. Т. е. в этом случае нельзя гарантировать, что косвенные измерения будут подчиняться именно нормальному распределению.

Эти особенности отражают серьезные проблемы практики эксперимента, связанные с априорной неопределенностью в отношении объектов мониторинга, которые являются сложными системами или киберфизическими системами.

Именно сложные системы определяют современные технологические прорывы и являются в настоящее ключевыми объектами экспериментов.

II. Имитационная модель закона распространения ошибок

Таким образом, практика показывает, что для широкого набора задач мониторинга сложных систем неприменима не только стандартная формула распространения ошибок (1), но и трактовка измеряемых величин с погрешностью, как случайных величин, подчиняющихся нормальному закону с заданными математическим ожиданием и дисперсией. Поэтому непараметрическая статистика [2] «подсказывает», что необходимо выбрать другое толкование погрешностей, которое являлось бы обобщением на произвольные распределения. Понятия математического ожидания и среднеквадратического отклонения можно естественным образом обобщить при помощи квантилей.

В качестве простейшего инженерного примера можно привести, например, квантиль уровня 0,5 (медиана) который определён для любого непрерывного распределения, а в случае нормального распределения совпадает с математическим ожиданием. Квантили 0,16 и 0,84 являются границами доверительного интервала с уровнем доверия 0,68, что в случае нормального распределения сводится к интервалу с границами $(\mu - \sigma, \mu + \sigma)$. Таким образом, для произвольных распределений мы будем использовать медиану в роли «обобщённого среднего», вместо математического ожидания, а разности квантиля уровня 0,84 с медианой $(0{,}84 - 0{,}5)$ и медианы с квантилем уровня 0,16 $(0{,}5 - 0{,}16)$ как погрешности сверху и снизу соответственно.

В данной работе предлагается непараметрический метод вычисления погрешностей прямых и косвенных измерений, позволяющий преодолеть описанные выше проблемы. Этот метод использует наработанный потенциал непараметрической статистики и современные вычислительные возможности IT технологий для имитационного статистического моделирования процессов косвенных измерений (использование метода Монте-Карло).

Идея заключается в следующем: поскольку для каждой входной измеренной величины мы имеем измеренную величину, которая описывается непрерывным распределением из широкого семейства экспоненциальных распределений и распределений с параметрами сдвига и масштаба. Означенное семейство охватывает практически все возможные на практике распределения и может быть описано на основе, предложенных выше, непараметрических квантильных оценок погрешностей. Кроме того эти распределения являются предельными распределениями не нормального типа, что и объясняет их широкое распространение в природе и технике [2].

Таким образом, означенные непараметрические оценки можно использовать как оценки погрешности прямых измерений в ситуации, когда вид распределения экспериментальных данных недостаточно известен. Т. е. оценивая параметры сдвига и масштаба реальных измерений можно оценивать погрешности прямых измерений.

Следующий этап это решение проблемы представления функции формирования косвенных измерений. Аналитика преобразования случайных величин (теория функционального преобразования совокупности случайных величин) представляет трудно разрешимую аналитическую задачу, поэтому в данном подходе предлагается воспользоваться методом Монте-Карло для генерирования модели прямых измерений с распределением из выше описанного класса, и определяемом на основе достаточно общих и физически очевидных свойств (например, непрерывность) физических величин объектов мониторинга сложных систем. Непараметрические оценки исходных прямых измерений используются для задания характеристик модельных распределений. Выбор функции распределения для моделирования прямых измерений из означенных семейств может быть обусловлен, либо физикой процессов в самом явлении, либо среди так называемых предельных (или устойчивых) распределений (например, распределения экстремальных значений, Коши и др.) [2, 3].

В практических приложениях часто возникает ситуация, когда измерения описываются несимметричными функциями распределения (логно-нормальное, гамма, F-распределения и др.). Здесь приходиться для описания погрешностей прямых измерений привлекать параметры асимметрии и эксцесса. На практике многие инженеры, экономисты, и физики говорят об «ассиметричных погрешностях» и предлагают эмпирические процедуры их оценки.

После генерации совокупности случайных величин, моделирующих прямые измерения, задаются функции их преобразования. Вид этих функций определяется особенностями изучаемых явлений и процессами их наблюдений. Затем происходит преобразование модельных прямых измерений в косвенные на основе физически обоснованных функций преобразования. Предлагаемый метод отражает и развивает идею непараметрического оценивания функционалов, предложенную Ф.П. Тарасенко [2].

На завершающем этапе моделирования с помощью непараметрических оценок (предложенных выше) измеряются (на основе модельных преобразований) модельные погрешности косвенные измерений.

Другими словами выбираем непрерывное распределение (из широкого распределений с параметрами сдвига и масштаба) случайной величины (в простейшем случае нормальное распределение) и генерируем выборку случайных величин размера N, подчиняющуюся распределению, которое моделирует распределение прямых измерений. То есть каждое из входных (прямых) измерений представляется в виде выборки размера N.

Тождественность с реальными измерениями обеспечивается за счет квантильных оценок прямых погрешностей измерений на основе реальных прямых измерений. Полученные модельные выборки прямых измерений поступают на модели формирования косвенных измерений. Затем на основе модельных косвенных измерений вычисляются непараметрические (квантильные) оценки косвенных погрешностей.

Например, по этой выборке можно определить квантили уровней 0,16, 0,5 и 0,84, чтобы оценить косвенное измерение и его погрешности.

Метод Монте-Карло и современные вычислительные возможности при достаточном объеме модельных выборок позволяет определять косвенные погрешности с

приемлемой точностью. Данный подход экспериментально проверен на реальном материале обработки терабайтных обзоров [4].

Следует отметить, что предлагаемый метод обладает рядом преимуществ.

- Во-первых, он прост в реализации и интерпретации результатов.
- Во-вторых, этот метод универсален в том смысле, что он применим к любого вида преобразованиям прямых измерений в косвенные измерения.
- В-третьих, в этом методе любые корреляции между случайными величинами учитываются при последующих расчётах автоматически, без необходимости рассчитывать производные.
- Единственным существенным недостатком является увеличение времени расчётов в N раз (в нашем случае N = 10000), но мы живём в такую эпоху, когда можем себе это позволить.

Данный подход реализован в виде программного продукта и экспериментально проверен на реальном материале обработки терабайтных обзоров [4]. Для решения поставленных задач был написан код на языке Python. Соответствующий репозиторий опубликован в открытом доступе по адресу https://github.com/Roustique/sngrb. Используемые библиотеки:

- NumPy – для поддержки массивов, в том числе многомерных, и функций, которые с ними работают.
- SciPy – использованы процедуры optimize.least_sqares – для нелинейной регрессии, stats.mstats.theilslopes – для линейной регрессии методом Тейла-Сена. Модуль stats – для использования реализованных интегральных и дифференциальных функций распределения, а также квантилей.
- Pandas – для работы с каталогами.
- Matplotpib, Seaborn, mpl_scatter_density – для построения изображений (графиков, диаграмм).
- Joblib – для параллелизации.
- Numba – для jit-компиляции.

Программный продукт обеспечивает возможность эмпирической оценки статистической достоверности измеряемых погрешностей косвенных измерений.

III. Заключение

Предлагаемый подход к вычислению погрешностей косвенных измерений развивает тенденцию многомерной обработки экспериментальных данных в условиях глубокой априорной неопределенности в отношении характеристик объектов мониторинга. В этом подходе измеряется совокупность различных физических или экономических величин и затем моделируется преобразование прямых погрешностей в косвенные (многомерная задача). Данный подход позволяет методами машинного эксперимента решать широкий спектр задач многомерной статистики и распознавания образов. Появляется возможность визуализации этих многомерных процессов средствами когнитивной машинной графики, что обеспечивает своевременное выявление аномальных феноменов в глубоком обучении нейронных сетей.

Так, например, предлагаемый метод имитационного моделирования процесса прямого распространения ошибок, может быть использован и для расчетов в нейронных сетях с обратным распространением ошибок.

Кроме того, подобный подход может способствовать решению фундаментальных проблем нейронных сетей – проблем извлечения правил из искусственных нейронных сетей [5, 6]. Так же моделирование формирования погрешностей косвенных измерений многомерных измерений может способствовать выяснению причин формирования аномальных погрешностей и байесов в многослойных нейронных сетях.

Список литературы

[1] Рекомендации по межгосударственной стандартизации. Государственная система обеспечения единства измерений. Метрология. Основные термины и определения. РМГ 2999.

[2] Тарасенко Ф.П. Непараметрическая статистика. Томск: Изд-во Томского университета, 1976, 202 с.

[3] Добровидов А.В., Кошкин Г.М. Непараметрическое оценивание сигналов. М:Наука.Физматлит. 1997. 336 с. ISBN 5-02-015217-X

[4] Lovyagin N.Yu., Gainutdinov R.I., Shirokov S.I., Gorokhov V.L. The Hubble Diagram: Jump from Supernovae to Gamma-ray Bursts // Universe. 2022, vol. 8(7). P. 344.

[5] Аверкин А.Н. Методы объяснимого искусственного интеллекта в работах Лотфи Заде // Материалы Научно-методического семинара «Искусственный интеллект в системах управления» 28 – 29 мая 2022 г. г. Дубна.

[6] Аверкин А.Н., Ярушев С.А. Обзор исследований в области разработки методов извлечения правил из искусственных нейронных сетей // Известия РАН. Теория и системы управления, 2021, № 6, стр. 106-121.