Метод непараметрической оценки закона распределения случайного параметра по малому числу наблюдений

Е.Б. Горбунова

Южный федеральный университет, Таганрог

Аннотация: В статье изложены результаты разработки и исследования метода проверки гипотез о виде функции плотности распределения случайной величины в условиях значительной априорной неопределенности. Полученные результаты, вопреки традиционному скептицизму в отношении обработки выборок данных, объемом порядка десяти значений, показывают потенциальную возможность повышения достоверности их классификации.
Ключевые слова: обработка статистической информации, малая выборка, численный метод, метод имитационного дополнения, статистический эксперимент, случайный процесс.

Введение

Нестационарные системы характеризуются быстрым изменением значений параметров, поэтому для осуществления их эффективного мониторинга представляется целесообразным использование методов статистического анализа случайных процессов, ориентированных на работу с малым числом наблюдений. В основу традиционных методов обработки статистической информации положена идея группировки данных (гистограммы, Критерий Пирсона и пр.), что при анализе выборок значительного объема позволяет добиться заданной достоверности оценок. Однако, как показано в работе [1], группировка наблюдений неизбежно связана с потерей информации, которую теоретически возможно извлечь из массива данных. Это говорит о том, что выборки большого объема содержат избыточную для достижения заданной точности оценок информацию. Исходя из этого, можно естественным образом определить понятие «малой выборки»: выборку следует считать «малой», если при ее обработке методами, основанными на группировке наблюдений, нельзя достичь заданной точности [1].

Таким образом, при работе с малыми выборками данных следует отказаться от группировки наблюдений и перейти к методам, основанным на использовании каждой отдельной реализации. В работе представлен метод имитационного дополнения малой выборки, основанный, во-первых, на идее аддитивной аппроксимации плотности распределения случайной величины симметричными вкладами [1,4,6], во-вторых, на использовании численных методов и возможности имитационного моделирования случайных процессов при помощи современных ЭВМ [3].

Метод имитационного дополнения

Суть метода имитационного дополнения состоит в генерации дополняющих массивов в окрестности каждого элемента исходной выборки, как показано на рис.1. Этот процесс логически близок к сглаживанию ступенчатой функции распределения и позволяет свести обработку малой выборки к существующим хорошо разработанным технологиям, таким как, например, критерий Пирсона, который, как известно, дает устойчивый результат при анализе выборок данных, объемом более пятидесяти значений [2].

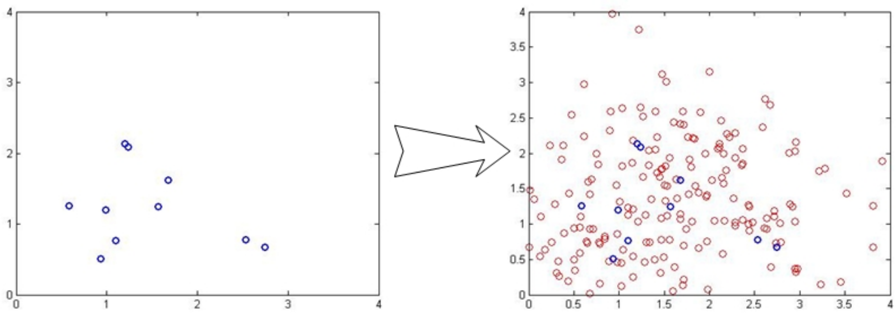

Рис. 1. – Имитационное дополнение малой выборки

Для исследования предлагаемого метода в программной среде MatLab был проведен следующий статистический эксперимент. Формировалась генеральная совокупность (ГС) путем преобразования массива данных, распределенных по равновероятному закону. Например, для получения распределения Рэлея:

$$y_i = \sigma \cdot \sqrt{-2\ln{(yu_i)}} \quad (1)$$

где yu_i – случайные числа, распределенные равномерно; σ - параметр распределения.

Генеральная совокупность проверялась на соответствие заданному закону распределения критерием согласия Пирсона. Если проверка давала значение χ^2, соответствующие вероятности менее 0.5, результаты эксперимента отбрасывались, поскольку принималась гипотеза о несостоятельности начальных условий [4]. Из генеральной совокупности извлекалась серия малых выборок путем формирования массива случайных номеров элементов ГС. Для анализа были выбраны следующие законы распределения: распределение Рэлея; Нормальное распределение; Логарифмическое нормально распределение; Экспоненциальное и Бета распределения.

Метод имитационного дополнения реализовывался следующим образом. При помощи встроенных средств MatLab генерировалась серия случайных величин y^d, математические ожидания которых совпадали с соответствующими элементами анализируемой выборки (назовем их вкладами по аналогии с методом аддитивной аппроксимации), а дисперсия вычислялась по формуле:

$$D^d = k(max - min) \quad (2)$$

где max и min – априорно известные границы диапазона изменения параметра, в эксперименте они брались равными наибольшему и наименьшему значениям ГС соответственно; k – коэффициент дисперсии вклада, $k \in (0; 0.5)$.

Расширенная выборка (РВ) формировалась в соответствии со следующим правилом:

$$Y = \cup_n y_n^d \tag{3}$$

Классификация осуществлялась следующим образом. По малым выборкам производилась оценка математического ожидания и дисперсии, в соответствии с которыми задавался ряд гипотетических распределений. Затем выборки (как малые, так и расширенные) проверились при помощи критерия согласия Пирсона на степень соответствия каждому из гипотетических распределений. Из полученных значений χ^2 строился вариационный ряд. Как истинная принималась гипотеза о распределении, давшем наименьшее значение χ^2 в этом ряду. Поскольку исходное распределение ГС известно, имелась возможность оценить число ошибок классификации. Следует отметить, что абсолютные значения χ^2 расширенных выборок значительно превышали значения, рассчитанные для необработанных выборок, однако как устойчивость, так и различимость результатов в этом случае была выше.

Совершенно очевиден тот факт, что достоверность классификации в значительной степени зависит от параметров вкладов. Для определения оптимальных значений n и k был проведен двухфакторный эксперимент, позволивший получить зависимости числа верных классификаций от этих параметров. Алгоритм данного эксперимента представлен на рис. 2.

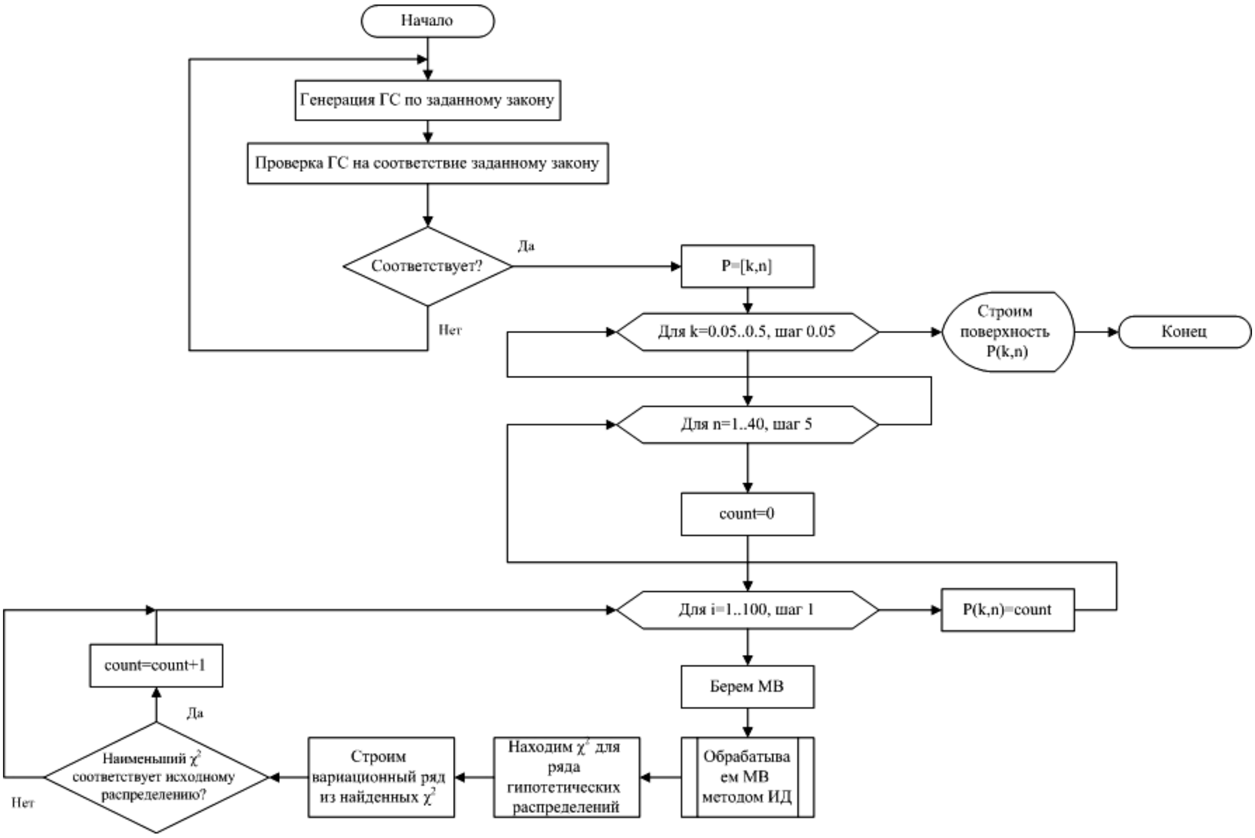

Рис. 2. – Алгоритм оценки зависимости эффективности метода от параметров вкладов.

Зависимости числа правильно классифицированных выборок от коэффициента дисперсии вклада k и числа элементов во вкладе n для различных законов распределения случайной величины представлены на рис. 3 – 5.

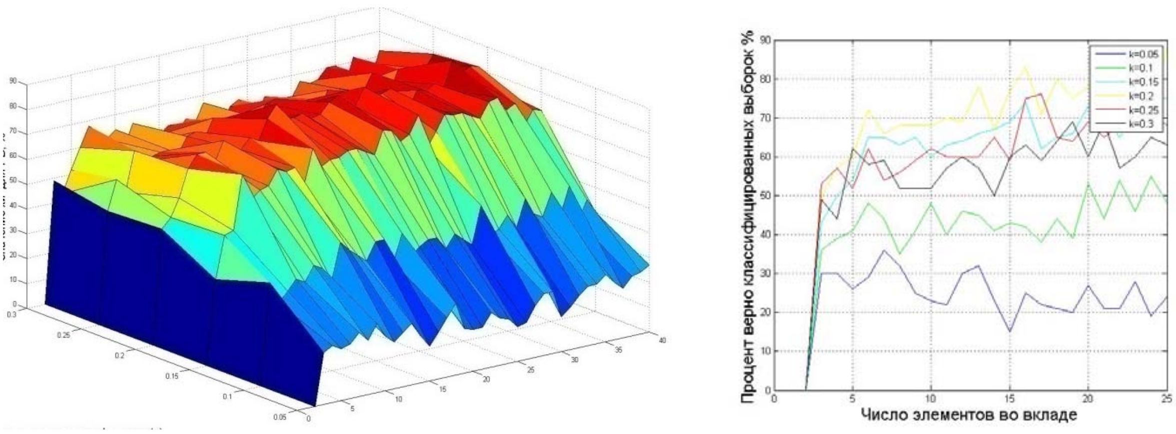

Рис. 3 – Зависимость числа правильно классифицированных выборок (из 100) от коэффициента дисперсии вклада *k* и числа элементов во вкладе *n* для распределения Рэлея.

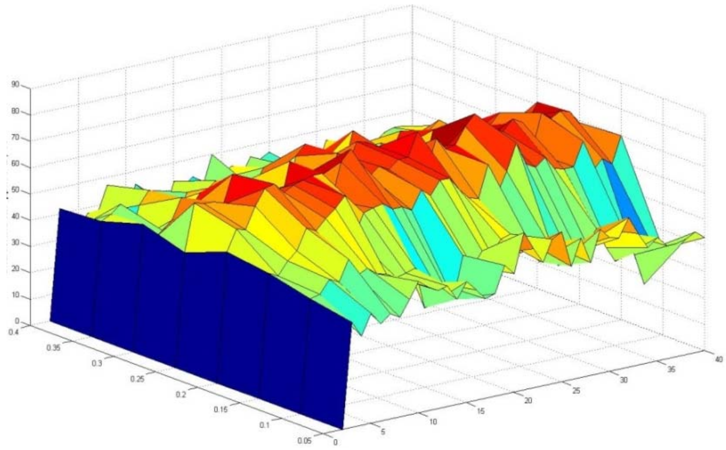

Рис. 4 – Зависимость числа правильно классифицированных выборок (из 100) от коэффициента дисперсии вклада *k* и числа элементов во вкладе *n* для Логарифмического нормального распределения.

Из графиков, показанных на рис. 3 – 4, видно, что для распределения Рэлея и Логарифмического нормального распределений число верно классифицированных выборок максимизируется при $n>10$ и $k=0.2$. Для нормального распределения оптимальные значения параметров другие: $k=0.1$, $n>10$, что следует из рис. 5.

Таким образом, при правильном выборе параметров вкладов количество верно классифицированных выборок может достигать 80% [10], однако сам факт зависимости оптимального значения коэффициента дисперсии вклада от вида плотности распределения исходной случайной величины, очевидно, требует дальнейшего исследования.

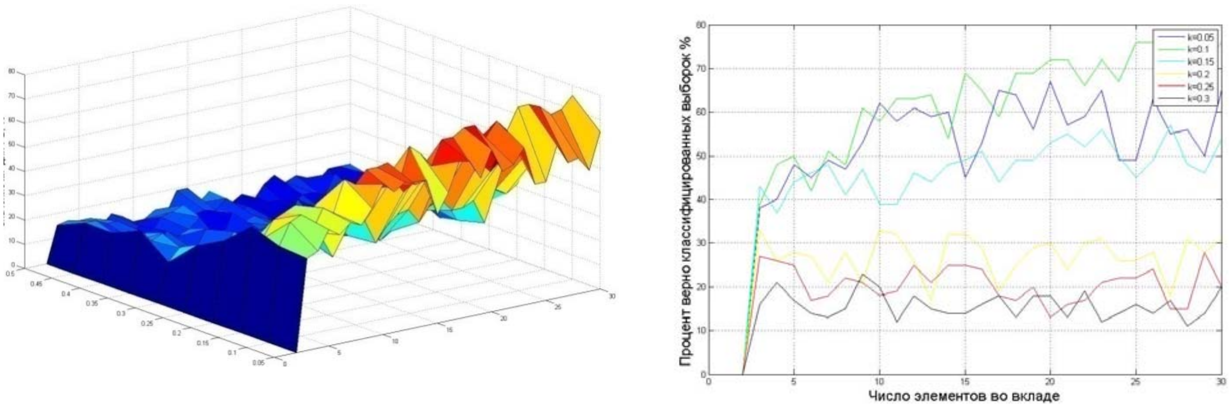

Рис. 5 – Зависимость числа правильно классифицированных выборок (из 100) от коэффициента дисперсии вклада k и числа элементов во вкладе n для Нормального распределения.

Результаты исследований, изложенные в данной статье, получены при финансовой поддержке Минобрнауки РФ в рамках реализации госзадания №213.01-11/2014-47 «Разработка систем диагностики состояния биологических и технических объектов с использованием алгоритмов анализа нестационарных сигналов».

Литература

1. Гаскаров Д.В., Шаповалов В.И. Малая выборка. М.: Статистика, 1978. 248 с.
2. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. М.: ФИЗМАТЛИТ, 2006. 816 с.

3. Диаконис П., Эфрон Б. Статистические методы с интенсивным использованием ЭВМ // Bootstrap. The private blog of Alexander Bulgakov URL: http://boot-strap.ru/ (дата обращения: 10.05.2014г.).
4. Algebraic laws for nondeterminism and concurrency. URL: citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.409.1770&rank=1
5. Жовинский А.Н., Жовинский В.Н. Инженерный экспресс-анализ случайных процессов. М: Энергия, 1979. 112 с.
6. Гузик В.Ф., Кидалов В.И., Самойленко А.П. Статистическая диагностика неравновесных объектов. СПб: Судостроение, 2009. 304 с.
7. Лапко А.В., Шарков Н.А. Непараметрические методы обнаружения закономерностей в условиях малых выборок. Приборостроение 2008. №8, Т.51., с. 62-67.
8. Демченко Д.Б., Касьянов В.Е. Оптимизационный метод статического расчета строительных конструкций с применением вероятностных законов с ограничениями. // Инженерный вестник Дона, 2013, №2 URL: ivdon.ru/magazine/archive/ n2y2013/1659
9. Ковалева А.В. Экономико-математическая модель оценки стратегического риска при выборе стратегии развития примышленного предприятия. // Инженерный вестник Дона, 2012, №1 URL: ivdon.ru/magazine/archive/ n1y2012/685
10. Flocks, herds, and schools: a distributed behavioral model. URL: citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.317.3619&rank=4

References

1. Gaskarov D.V., Shapovalov V.I. Malaja vyborka. M.: Statistika, 1978. 248 s.
2. Kobzar' A.I. Prikladnaja matematicheskaja statistika. Dlja inzhenerov i nauchnyh rabotnikov. M.: FIZMATLIT, 2006. 816 s.

3. Diakonis P., Jefron B. Statisticheskie metody s intensivnym ispol'zovaniem JeVM // Bootstrap. The private blog of Alexander Bulgakov URL: http://boot-strap.ru/ (data obrashhenija: 10.05.2014g.)
4. Algebraic laws for nondeterminism and concurrency URL: citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.409.1770&rank=1
5. Zhovinskij A.N., Zhovinskij V.N. Inzhenernyj jekspress-analiz sluchajnyh processov. M: Jenergija, 1979. 112 s.
6. Guzik V.F., Kidalov V.I., Samojlenko A.P. Statisticheskaja diagnostika neravnovesnyh ob#ektov. SPb: Sudostroenie, 2009. 304 s.
7. Lapko A.V., Sharkov N.A. Neparametricheskie metody obnaruzhenija zakonomernostej v uslovijah malyh vyborok. Priborostroenie 2008. №8, T.51., s. 62-67.
8. Demchenko D.B., Kas'janov V.E. Inženernyj vestnik Dona (Rus), 2013, №2 URL: ivdon.ru/magazine/archive/ n2y2013/1659
9. Kovaleva A.V. Inženernyj vestnik Dona (Rus), 2012, №1 URL: ivdon.ru/magazine/archive/ n1y2012/685
10. Flocks, herds, and schools: a distributed behavioral model. URL: citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.317.3619&rank=4