

УДК 004.421.6: 519.23

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ СТАТИСТИЧЕСКОГО ОЦЕНИВАНИЯ ЭНТРОПИИ

Р.А. КАПУСТО, В.Ю. ПАЛУХА

Белорусский государственный университет, г. Минск, Республика Беларусь

Введение. Одним из подходов к оценке качества генераторов случайных и псевдослучайных последовательностей является статистическое оценивание энтропии и сравнение полученной оценки с ожидаемым значением для равномерно распределенной случайной последовательности (РРСП). Данная работа посвящена сравнительному анализу методов статистического оценивания энтропии.

1. Обзор методов оценки энтропии. Пусть имеется случайная последовательность $\{x_i : i = 1, \dots, n\}$ из распределения вероятностей $\{p_k : k = 1, \dots, N\}$. Энтропия Шеннона данной последовательности вычисляется по формуле [1]:

$$H(P) = - \sum_{k=1}^N p_k \log p_k. \quad (1)$$

Поскольку истинные значения вероятностей неизвестны, мы можем лишь вычислить оценку энтропии последовательности. В работе рассмотрены несколько методов: подстановочный метод [2], метод Миллера-Мэдоу [3], байесовский метод [3], метод Грассбергера [4] и метод усадки Джеймса-Штейна [3].

Одним из подходов к статистическому оцениванию энтропии является построение частотных оценок вероятностей $\{\hat{p}_k\}$ и подстановка полученных оценок в функционал энтропии вместо истинных значений вероятностей $\{p_k\}$. По подстановочному методу построение частотных оценок для вероятностей производится по следующим формулам:

$$\hat{p}_k = \frac{v_k}{n}, \quad v_k = \sum_{i=1}^n I\{x_i = \omega_k\}, \quad I\{x_i = \omega_k\} = \begin{cases} 1, & x_i = \omega_k; \\ 0, & x_i \neq \omega_k. \end{cases} \quad (2)$$

Методом Миллера-Мэдоу называют подстановочный метод, скорректированный с помощью константы,

$$\hat{H}_{MM} = \hat{H}_{plug-in} + \frac{\hat{N} - 1}{2n} \log e, \quad (3)$$

где $\hat{H}_{plug-in}$ – оценка, полученная подстановочным методом, $\hat{N}$ – оценка количества исходов с ненулевыми вероятностями.

Метод Байеса основан на корректировке оценок вероятностей и является еще одной модификацией подстановочного метода. Метод повторяет шаги плаг-ин оценки, только вместо формулы (2) используется следующая оценка:

$$\hat{p}_k^{Bayes} = \frac{v_k + a_k}{n + A}, \quad (4)$$

где a_k – поправочные коэффициенты, $A = \sum_{k=1}^N a_k$, $k = 1, \dots, N$.

Таким образом, оценка принимает следующий вид:

$$\hat{H}^{Bayes} = - \sum_{k=1}^N \hat{p}_k^{Bayes} \log \hat{p}_k^{Bayes}. \quad (5)$$

Метод Грассбергера также построен на основе подстановочного метода. Формула оценки следующая:

$$\hat{H}_G = \log n - \frac{1}{n} \sum_{k=1}^N v_k G_{v_k}, \quad G_k = \psi(k) + (-1)^k \int_0^1 \frac{x^{k-1}}{x+1} dx, \quad (6)$$

где $\psi(x) = \frac{d \ln \Gamma(x)}{dx}$ – дигамма функция, v_k вычисляется по формуле (2).

Усадка Джеймса-Штейна основана на усреднении двух абсолютно разных моделей: многомерной модели с низким смещением и высоким отклонением и модели более низкой размерности с большим смещением, но меньшей дисперсией. Интенсивность усреднения определяется относительным весом входящих в состав моделей. Коэффициент λ для выпуклой комбинации

$$\hat{p}_k^{Shrink} = \lambda t_k + (1-\lambda) \hat{p}_k^{ML} \quad (7)$$

является коэффициентом интенсивности усадки. Он принимает значения от 0 (нет усадки) до 1 (полная усадка). В формуле (7) $\hat{p}_k^{ML}$ – оценки вероятностей подстановочного метода, вычисленные по формулам (2), t_k – цель усадки.

Формула для вычисления коэффициента интенсивности усадки принимает вид

$$\hat{\lambda} = \frac{1 - \sum_{k=1}^N \left(\hat{p}_k^{ML} \right)^2}{(n-1) \sum_{k=1}^N \left(t_k - \hat{p}_k^{ML} \right)^2}. \quad (8)$$

Оценка энтропии по методу усадки Джеймса-Штейна производится по формуле

$$\hat{H}^{Shrink} = - \sum_{k=1}^N \hat{p}_k^{Shrink} \log \hat{p}_k^{Shrink}. \quad (9)$$

Для подстановочного метода в статье [5] приведено математическое ожидание оценки энтропии (используется натуральный логарифм):

$$E \left\{ \hat{H}_{plug-in} \right\} = \ln n - e^{-\lambda} \sum_{k=1}^{+\infty} \frac{\ln(k+1) \lambda^k}{k!}, \quad \lambda = \frac{n}{N}. \quad (10)$$

Для этого метода вычислено математическое ожидание оценки энтропии в случае, если наблюдаемая последовательность является РРСП. Полученное значение использовалось в дальнейшем исследовании.

2. Сравнительный анализ. Анализ проводился на основе следующих показателей: среднее значение, среднеквадратическая ошибка, выборочная дисперсия. Данные принадлежали одному из двух алфавитов: бинарный (мощности 2) и байтовый (мощности 256). Для каждого алфавита получено по тысяче последовательностей из дискретного равномерного распределения размерами 10, 30, 50, 100, 300, 500, 1000, 3000 и 5000. В качестве генератора использовался стандартный генератор псевдослучайных символов языка программирования Python.

На первом этапе исследования были вычислены оценки энтропии рассмотренными методами, а также найдены средние значения указанных выше показателей. Далее цветами обозначены следующие методы: красный – подстановочный метод, желтый – метод Миллера-Мэдоу, зеленый – байесовская оценка, синий – метод Грассбергера, фиолетовый – метод усадки Джеймса-Штейна. На рисунках для построения графика выбраны наиболее информативные значения длин последовательностей (длины, для которых линии методов накладываются друг на друга, не включены).

На рисунке 1, где приведен график зависимости среднеквадратичного отклонения от длины последовательности для байтовой последовательности, показано, что значения, полученные по методу усадки Джеймса-Штейна, меньше среднего соответствующего параметра остальных методов. Это указывает на то, что значения энтропии, полученные методом усадки, ближе к эталонному значению энтропии. Смещение оценки энтропии, полученное по ме-

тоду Миллера-Мэдоу, меньше смещения соответствующей подстановочной оценки на всем рассматриваемом промежутке – как и требует построение метода.

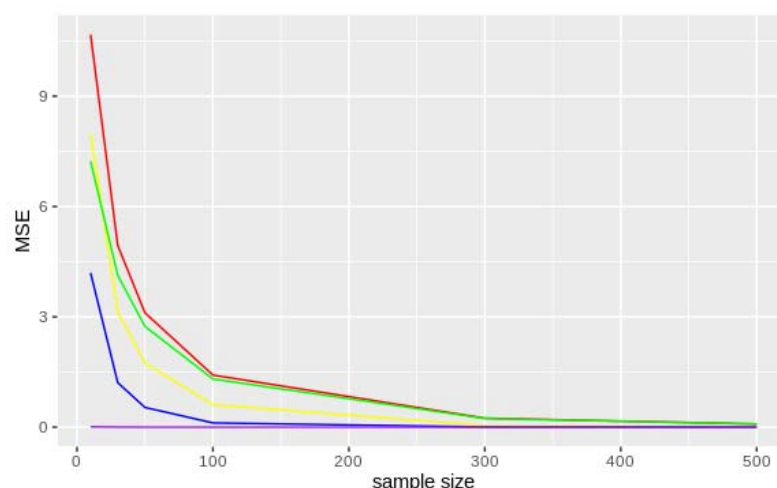


Рис. 1. Среднеквадратическое отклонение на последовательности байтов (равномерное распределение)

На рисунке 2 приведен аналогичный график для битовых последовательностей. Самое высокое среднеквадратическое отклонение у метода усадки Джеймса-Штейна.

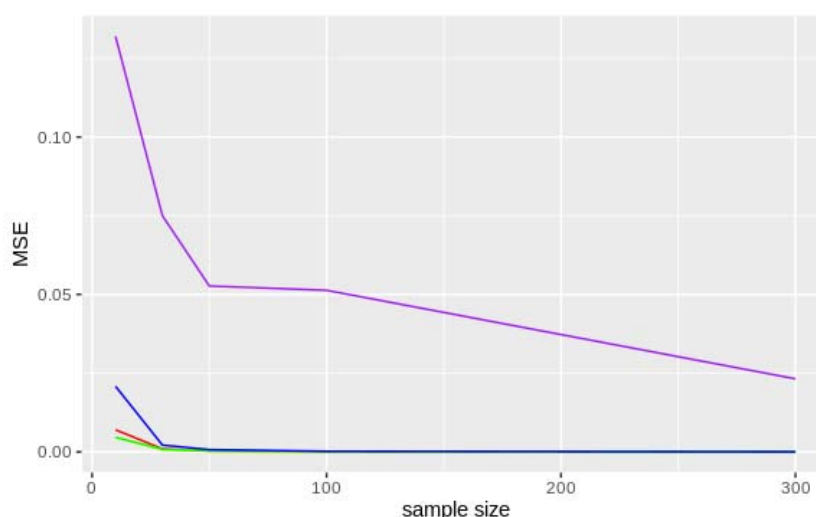


Рис. 2. Среднеквадратическое отклонение на последовательности битов (равномерное распределение)

На втором этапе исследования входные данные были модифицированы следующим образом: каждый десятый символ в последовательности битов был заменен нулевым, каждый сотый в последовательности байт – нулевым байтом. Целью описанного изменения было намеренно понизить энтропию входных данных, чтобы проследить реакцию методов на подобные изменения. Полагается, что чем больше отличаются оценки энтропии равномерно распределенных и модифицированных последовательностей, тем точнее мы сможем отличить последовательности таких генераторов в случае, когда нам неизвестно, продолжает генератор порождать равномерно распределенную случайную последовательность либо по каким-то причинам (поломка, подмена и т. п.) выходная последовательность не является равномерно распределенной. Для полученных последовательностей вычислены средние значения энтропии и другие показатели, как и для первого этапа исследования.

На рисунке 3 представлен график, отражающий разницу между средним значением энтропии для байтов на первом и втором этапе исследования. Для метода усадки Джеймса-Штейна разница наименьшая, что означает, что метод хуже всего откликается на понижение энтропии.

Рисунок 4 отражает аналогичную разницу для последовательностей битов. Так как на первом этапе метод усадки Джеймса-Штейна показал худший результат, этот метод не взят в рассмотрение на рисунке 4.

Метод Грассбергера имеет меньшую разницу из всех методов. Остальные методы имеют приблизительно одинаковые показатели.

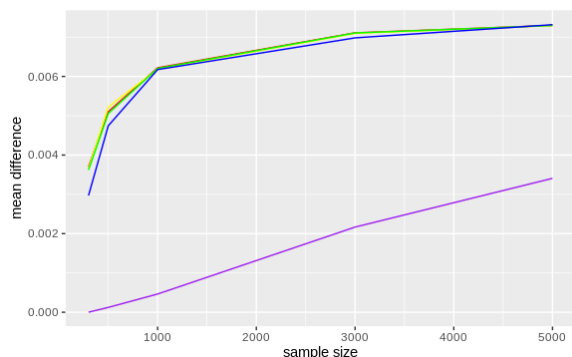


Рис. 3. Разность среднего значения энтропии на последовательности байтов

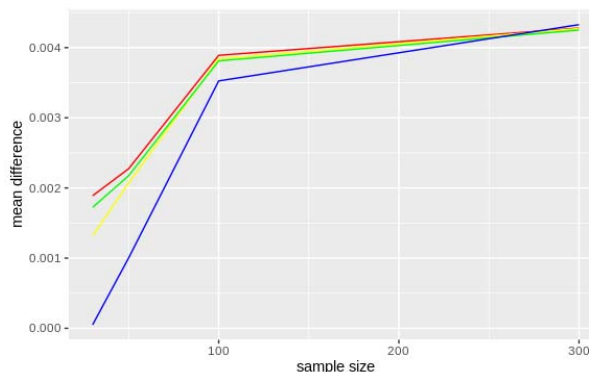


Рис. 4. Разность среднего значения энтропии на последовательности битов

Заключение. По результатам проведенных исследований можно сделать следующие выводы о методах оценки энтропии Шеннона.

1. Плаг-ин оценка: стабильна, средние относительно других методов показатели смещения и дисперсии. Базовый метод.
2. Метод Миллера-Мэдоу: показатели смещения лучше, но уступает в дисперсии плаг-ин оценке.
3. Байесовская оценка: результаты, как и у плаг-ин оценки.
4. Метод Грассбергера: малое отклонение, но большая дисперсия. Метод хуже других проявил себя на втором этапе исследований по распознаванию модифицированных последовательностей.
5. Метод усадки Джеймса-Штейна: плохие результаты для битовых последовательностей. Так как на втором этапе исследований метод плохо себя проявил, его применимость остается под вопросом.

Список литературы

1. Shannon, C. E. A mathematical theory of communication / C. E. Shannon // Bell. System Tech. – 1948. – J. 21. – P. 379–423.
2. Башарин, Г. П. О статистической оценке энтропии последовательности независимых случайных величин / Г. П. Башарин // Теория вероятн. и ее примен. – 1959. – Т. 4, вып. 3. – С. 361–364.
3. Hausser, J. Entropy Interference and the James-Stein Estimator, With Application to Nonlinear Gene Association Networks / J. Hausser, K. Strimmer // Journal of Machine Learning Research. – 2009. – J. 10. – P. 1469–1484.
4. Grassberger, P. Entropy Estimates from Insufficient Sampling [Electronic resource] / P. Grassberger. – Режим доступа : <https://arxiv.org/pdf/physics/0307138.pdf>. – Дата доступа : 31.03.2021.
5. Палуха, В. Ю. Статистические тесты на основе оценок энтропии для проверки гипотез о равномерном распределении случайной последовательности / В. Ю. Палуха // Весці НАН Беларусі. Серыя фізіка-матэматычных навук. – 2017. – № 1. – С. 79–88.