Работоспособность непараметрического критерия Вилкоксона при решении задач с особенностями в выборках

А. Е. Коченгин
НИУ МЭИ
Москва
kochenginalexey@gmail.com

Г. Хрисостому
университет им. Фредерика
Кипр
Chrysostomou_George@yahoo.com

В. А. Шихин
НИУ МЭИ
Москва
ShikhinVA@mpei.ru

Аннотация. **Предлагается подход к повышению разрешимости задач дискриминации данных посредством непараметрического критерия знаковых рангов Вилкоксона при принятии решений в ситуациях с особенностями в сравниваемых выборках данных типа "смещение по вертикали" и "масштаб". Модифицированный критерий Вилкоксона рассматривается в качестве инструмента выявления отклонений при ведении технологического процесса, которые не превышают пороговые значения, но оказывают существенное влияние на технико-экономические показатели. Предложенный метод выявления критических событий с применением модифицированного критерия Вилкоксона позволяет определять как факт наличия события, так и производить их распознавание. Доказывается, что мощность предложенного модифицированного критерия Вилкоксона не ухудшается по сравнению с классическим критерием Вилкоксона. Произведенная обработка учетных данных на примере электропотребления одного из нефтехимических промышленных предприятий позволила протестировать возможность применения модифицированного критерия Вилкоксона в составе вычислительной процедуры разработанного алгоритма выявления критических событий.**

Ключевые слова: непараметрический критерий; технологическое событие; статистические методы; профиль нагрузки

I. Введение

Проблема пропуска важной информации, не нарушающей в целом технологический процесс, однако оказывающей существенное влияние на оптимальное функционирование объекта управления и контроля является предметом многих исследований. Один из используемых при этом подходов сводится к сравнительному анализу соответствующих выборок данных.

При решении задач дискриминации двух выборок перед исследователем встает вопрос выбора критерия. Наиболее распространённым является использование статистических методов. Однако для выборок с небольшим количеством значений (до 50), наиболее известные параметрические подходы не работают. В таких случаях обычно применяют непараметрические критерии.

Особенность постановки задачи тестирования двух выборок на предмет принадлежности к единой выборке и отсутствию существенных отклонений в силу наличия критических событий в рассматриваемом случае сравнения профилей электропотребления состоит в том, что сравниваются две упорядоченные по времени выборки Y_1 и Y_2 поскольку они состоят из почасовых измерений электроэнергии.

Предполагается, что профиль нагрузки, по которому формируется выборка Y_1, соответствует интервалу времени, на котором гарантированно отсутствовали критические события, а Y_2 является выборкой данных по исследуемому на предмет факта наличия или отсутствия критических событий по другому интервалу времени.

II. Предлагаемый модернизированный критерий Вилкоксона

В практике применения критерия знаковых рангов Вилкоксона для выявления критических событий возникает ряд трудностей, связанных с особенностями, проявляющимися при попарном сравнении элементов классифицируемых выборок. Это так называемые альтернативы "масштаба" и "смещения по вертикали"**.**

Неучет данных обстоятельств может приводить к качественно неверному результату. Рассмотрим задачу выработки такого универсального приема обработки статистических данных, который бы позволил учесть указанные нарушения, а также исследуем вопросы,

связанные с изменением мощности усовершенствованного критерия.

Обойти указанные трудности предлагается за счет учета дополнительной априорной информации об известных (обычно) требованиях к точностным характеристикам данных. При учете электроэнергии точностные характеристики данных непосредственно зависят от классов точности применяемых электросчетчиков, а также от утвержденной в установленном порядке *Методики выполнения измерений* [1].

Допустим, что в соответствии с требованиями к точностным характеристикам в конкретном исследовании на исследуемую выборку Y_2 наложены поэлементно ограничения:

$$\begin{aligned} &\left|Y_{1i} - Y_{2i}\right| \le \sigma_0;\ for \forall i, i - \text{int}\, eger \\ &Y_{1i}, Y_{2i} \ge 0 \end{aligned} \qquad (1)$$

В выражении (1) положительная константа σ_0 задается как известная мера точности полученных измерительных данных.

Полагаем, что на основе априорной информации практически всегда может быть выбрана среди архива профилей нагрузки такая выборка Y_1 с гарантированным отсутствием критических событий и которая заведомо удовлетворяет точностным требованиям к отражению технологического процесса $\sigma_1 \le \sigma_0$, где σ_1 – медиана абсолютного отклонения данных по Y_1.

Пусть δ_{1i} и δ_{2i} представляют собой отклонения наблюдений Y_{1i} и Y_{2i} от гипотетических (без ошибок измерения) истинных значений элементов Y_{0i}, $i = \overline{1,N}$ эталонной выборки Y_0:

$$\begin{aligned} &\delta_{1i} = \left|Y_{1i} - Y_{0i}\right| \to \sigma_1 = Ex\left[\delta_1\right] \\ &\delta_{2i} = \left|y_{2i} - y_{0i}\right| \to \sigma_2 = Ex\left[\delta_2\right] \\ &\delta_i = \left|y_{1i} - y_{2i}\right| \to \sigma = Ex\left[\delta\right] \\ &\sigma_2 \ge \sigma_1 \,\&\, \sigma_1 \le \sigma_0 \end{aligned} \qquad (2)$$

Анализируя традиционную процедуру метода Вилкоксона, можно сделать вывод, что для самой неблагоприятной ситуации взаимного расположения численных значений наблюдений по сравниваемым Y_1 и Y_2 будем иметь:

$$\sigma \le \left(\sigma_1 + \sigma_2\right) \Rightarrow \sigma = \sigma_1 + \sigma_2 \qquad (3)$$

Для отмеченных выше особых ситуаций предлагается процедуру критерия Вилкоксона применить не к первичным разностям наблюдений δ_i, а проверять гипотезу об одностороннем расположении матожидания $\sigma = Ex[\delta]$ относительно заданной константы σ_0. Другими словами, матожидание от ошибок измерения должно быть меньше или равно наложенным на точностные характеристики ограничениям, что можно сформулировать в виде проверки гипотезы *H1:*

$$H1:\ Ex[\delta] < \sigma_0 \qquad (4)$$

Однако при этом встает вопрос, как скажется замена тестируемой гипотезы *H* на *H1* при вынесении окончательного решения о возможном наличии критического события. Для ответа на этот вопрос предлагается сформулировать и доказать следующую теорему.

Теорема 1. Положительное решение по гипотезе *H1* об одностороннем расположении матожидания разностей δ_i относительно заданного уровня точности σ_0: $H1: Ex[\delta] < \sigma_0$ вместо проверки исходной гипотезы *H* о принадлежности двух выборок Y_1 и Y_2 одной генеральной совокупности Y_0: $H: Y_1 \in Y_0$ & $Y_2 \in Y_0$ является лишь достаточным условием достоверности исходной гипотезы *H*.

Доказательство. Для доказательства Теоремы 1 необходимо проверить следующие два положения, соответствующие ситуациям отсутствия или наличия критических отклонений во второй из двух сравниваемых выборок Y_1, Y_2:

а). Верно ли, что если $Ex[\delta] < \sigma_o$, *то* $\sigma_2 < \sigma_o$?

б). Следует ли из того, что $Ex[\delta] > \sigma_o$, $\Rightarrow \sigma_2 > \sigma_o$?

При этом будем исходить из самой неблагоприятной ситуации (3) относительно взаимного расположения наблюдений.

Анализируя (а) с учетом (2), (3) с достаточной очевидностью получаем

$$Ex[\delta] = (\sigma_1 + \sigma_2) < \sigma_o \Rightarrow \sigma_2 < (\sigma_o - \sigma_1) \qquad (5)$$

Поскольку известно, что $\sigma_1 < \sigma_o$, следовательно, $\sigma_2 < \sigma_o$, что и требовалось доказать.

Анализируя (б) с учетом (2), (3) имеем:

$$Ex[\delta] = (\sigma_1 + \sigma_2) > \sigma_0 \Rightarrow \sigma_2 > (\sigma_0 - \sigma_1) \qquad (6)$$

Поскольку известно, что $\sigma_1 < \sigma_o$, следовательно, σ_2 не обязательно больше σ_o. Каждая из σ_1 и σ_2 может быть меньше σ_o, однако при этом их сумма может быть больше σ_o. Следовательно, положение (б) выполняется не всегда.

Таким образом, достаточные условия Теоремы 1 можно считать доказанными. Для частного случая из Теоремы 1 можно вывести очевидное следствие.

Следствие 1. Если предположить, что одна из выборок, например, Y_1 является эталонной, т.е. $\sigma_1 \equiv 0$, то положения (а) и (б) выполняются автоматически:

$$Ex[\delta] = \sigma_2, \& Ex[\delta] < \sigma_0 \Rightarrow \sigma_2 < \sigma_0 \qquad (7)$$

$$Ex[\delta] = \sigma_2, \& Ex[\delta] > \sigma_0 \Rightarrow \sigma_2 > \sigma_0 \qquad (8)$$

и полученные условия являются необходимыми и достаточными для принятия гипотезы (4).

Таким образом, для повышения мощности критерия Вилкоксона, предлагается модифицировать алгоритм традиционного критерия Вилкоксона**.** В результате имеем:

Шаг 1: Выдвижение гипотезы

Шаг 2: Вычисляются разности $\delta_i, i=\overline{1,N}$ из N пар наблюдений $(Y_{1\,1}, Y_{2\,1}), (Y_{1\,2}, Y_{2\,2}),..., (Y_{1\,N}, Y_{2\,N})$: $\delta_i=Y_1 - Y_2$, $i=\overline{1,N}$

Шаг 3: Вычисляются разности $\delta_i - \sigma_0, i=\overline{1,N}$. Значение σ_o задается на основе априорной информации.

Шаг 4: Вычисленные разности $\delta_i - \sigma_0, i=\overline{1,N}$ упорядочиваются по абсолютной величине в виде вариационного ряда $\delta(1)-\sigma_o, \delta(2)-\sigma_o,..., \delta(N)-\sigma_o$ и каждой разности в порядке возрастания присваивается соответствующий ранг R_i – целое положительное число: $R_i=\overline{1,N}, i=\overline{1,N}$.

Шаг 5: Каждому рангу $R_i=\overline{1,N}$ приписывается знак соответствующей разности $(Y_{1i} - Y_{2i} - \sigma_o)$, $i=\overline{1,N}$ и вычисляется сумма положительных рангов T_{N+}.

Шаг 6: Вычисленное значение T_{N+}. сравнивается с критическим значением критерия $A[\alpha,N]$, которое определяется из статистических таблиц [3] в соответствии с заданным уровнем значимости α и числом сравниваемых пар N.

III. Исследование мощности предложенного модернизированного критерия Вилкоксона

Рассматриваемый в данной работе критерий Вилкоксона относится к классу ранговых критериев, т.е. опирающихся на ранговую статистику $T=t(R)$, и принадлежит подклассу линейных критериев. Линейная ранговая статистика может быть представлена в общем случае в виде:

$$T=\sum_{i=1}^{N} a(i,R_i) \qquad (9)$$

где $\{a(i,j)\}$ есть произвольная матрица размером $[N\times N]$. В данном случае рассматриваем критерий знаковых рангов Вилкоксона, опирающийся на простую линейную статистику

$$T=\sum_{i=1}^{N} R_i \qquad (10)$$

Требуется установить, как повлияет предложенная выше модификация процедуры критерия на значение дисперсии σ^2, которая в свою очередь определяет значение мощности критерия дискриминации, а в приведенном случае использования линейного критерия – данная зависимость однозначна.

Предполагаем, что используемая в исследовании априорная информация (предпосылки) включает:

1. Наблюдения по показателям дискриминации принимают только неотрицательные значения: $y_i \geq 0$, $i=\overline{1,N}$. Считаем, что это условие всегда выполняется или может быть обеспечено без особых затруднений, учитывая, что данные в рассматриваемом случае представляют собой показания потребленной энергии.

2. Всегда может быть указана "эталонная выборка". Это могут быть результаты экспертных оценок архивов профилей нагрузки. Знание значения показателя дискриминации из эталонной выборки позволяет перейти от заданного безразмерного значения константы $\sigma_o\%$ к заданию ее абсолютного значения σ_o.

Начнем рассмотрение возможных особенностей в данных для случая "смещение по вертикали". Элементы выборки Y_1 на рис. 1 помечены как "♦", а выборки Y_2 как "■".

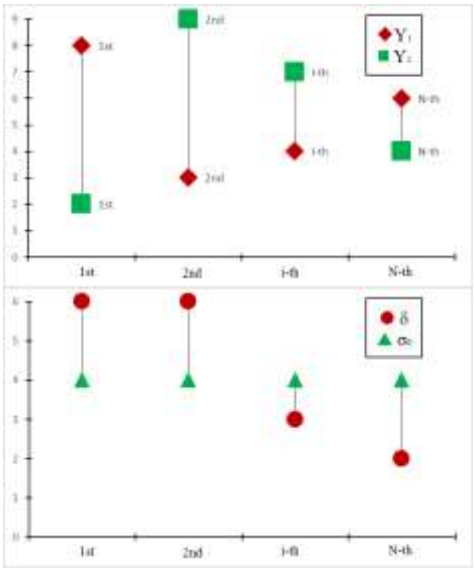

Рис. 1. Случай особенностей в данных типа "смещение по вертикали"

Из приведенной на рис. 1 графической интерпретации ясно, что смещение разностей $(\delta_i-\sigma_O)$, вычисляемых согласно модернизированному критерию приводит к уменьшению дисперсии $\sigma^2[\delta-\sigma_O]$ по сравнению с дисперсией $\sigma^2[\delta]$, имеющей место в традиционном алгоритме. Заметим, что это справедливо только при правильном выборе σ_O, что является особой задачей исследования.

Рассмотрим случай особенности типа "масштаб", когда большим по абсолютной величине, но разнознаковым разностям $\delta_i = y1_i - y2_i, i=\overline{1,N}$ сопутствует ситуация $T_{N+} \cong T_{N-}$ при которой критерий Вилкоксона в

традиционной форме неработоспособен. На рис. 2 учтена особенность взятия абсолютного значения от разностей δ_i.

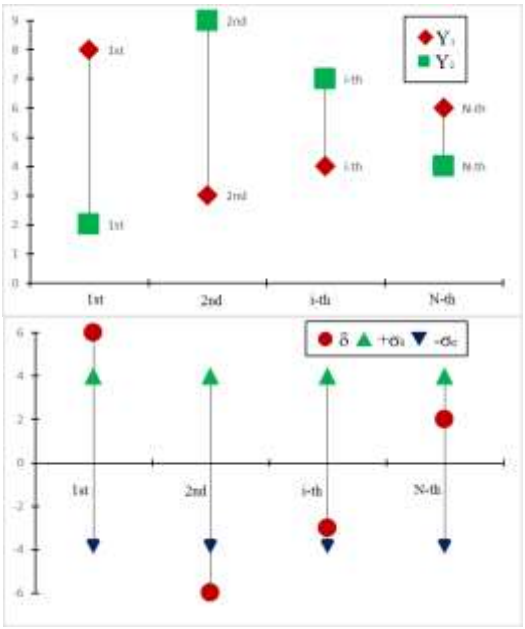

Рис. 2. Случай особенностей в данных типа "масштаб"

Из приведенных рисунков видно, что значение дисперсии $\sigma^2[\delta\text{-}\sigma_O]$, рассчитываемое в модернизированной процедуре по отношению к $(+\sigma_O)$ и $(-\sigma_O)$, уменьшается в сравнении с $\sigma^2[\delta]$.

Таким образом, доказано, что предложенное изменение традиционной процедуры критерия знаковых рангов Вилкоксона приводит к повышению мощности критерия и к приданию ему способности служить формализованным инструментом для принятия решений в ситуациях с нарушениями в данных типа "смещение по вертикали" и "масштаб".

IV. Выводы

Предложенная процедура модифицированного критерия Вилкоксона за счет учета доступной дополнительной априорной информации об известных требованиях к точностным характеристикам данных позволяет повысить разрешающую способность критерия знаковых рангов Вилкоксона.

Доказано, что мощность модифицированного критерия Вилкоксона по сравнению с традиционной процедурой метода в общем случае не ухудшается, а при особенностях в сравниваемых данных типа "смещение по вертикали", "масштаб" мощность критерия повышается.

Список литературы

[1] Spiering T. Energy efficiency benchmarking for injection moulding processes // Robotics and Computer-Integrated Manufacturing. 2015. №36.P.45–59.

[2] Кобзарь А.И. Прикладная математическая статистика. Физматлит: 2006. С.457-458.

[3] Owen D.B. Handbook of statistical tables. Adisson-wesley publishing company.:0. 1962. P.580.

[4] Крамер Г. Математические методы статистики, пер. с англ., 2 изд., М., 1975.

[5] Коченгин А.Е., Шихин В.А., Павлюк Г.П. Выявление и идентификация значимых технологических событий при анализе профиля электроснабжения промышленного предприятия // Автоматизация в промышленности 2019. №1. С.25-31.

[6] Орлов А.И. Современная прикладная статистика // Заводская лаборатория. Диагностика материалов. 1998. Т.64. №3. С. 52-60.

[7] Горский В.Г., Орлов А.И. Математические методы исследования: итоги и перспективы // Заводская лаборатория. Диагностика материалов. 2002. Т.68. №1. С.108-112.

[8] ГОСТ Р 50.1.037-2002. Рекомендации по стандартизации. Прикладная статистика: Правила проверки согласия опытного распределения с теоретическим. Часть II: Непараметрические критерии. Госстандарт РФ: 2002.С.66.

[9] Ассоциация НП Совет рынка - Приложение № 11.1 к Положению о порядке получения статуса субъекта оптового рынка и ведения реестра субъектов оптового рынка [Электронный ресурс], URL: https://www.np-sr.ru/sites/default/files/sr_regulation/reglaments/SR_0V048679/r1_1_pril_11_1_01092016_29082016.pdf

[10] Орлов А.И. Точки роста статистических методов // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета (Научный журнал КубГАУ) [Электронный ресурс]. Краснодар: КубГАУ, 2014. №09(103). С. 136–162. IDA [article ID]: 1031409011. Режим доступа: http://ej.kubagro.ru/2014/09/pdf/11.pdf

[11] Орлов А.И. Часто ли распределение результатов наблюдений является нормальным? // Заводская лаборатория. Диагностика материалов. 1991. Т.57. №7. С.64-66.

[12] Орлов А.И. Неустойчивость параметрических методов отбраковки резко выделяющихся наблюдений. // Заводская лаборатория. Диагностика материалов. 1992. Т.58. №7. С.40-42.

[13] Teiwes H Energy load profile analysis on machine level //CIRP. 2018. №69. P.271–276.

[14] Kang M. S Load profile synthesis and wind power generation prediction for an isolated power system // IEEE. 2007. P. 1459 – 1464.

[15] Hogg R.V Probability and Statistical Inference. Prentice Hall. 2006. P. 557.