ИСПОЛЬЗОВАНИЕ СТАТИСТИЧЕСКИХ КРИТЕРИЕВ ДЛЯ ОЦЕНКИ КАЧЕСТВА ДАННЫХ (НА ПРИМЕРЕ ДАННЫХ ПО СВОЙСТВАМ НЕОРГАНИЧЕСКИХ ВЕЩЕСТВ)

И.Д. Тарасенко[@], В.А. Дударев

¹Московский технологический университет (Институт тонких химических технологий),
Москва 119571, Россия

[@]Автор для переписки, e-mail: berkut257@gmail.com

В статье было проведено сравнение трех статистических критериев на основе выборки химических данных. При сравнении критериев приняли уровень значимости равным 0.05, поскольку это значение чаще всего используется в технических измерениях. С помощью статистических критериев была проведена оценка одностороннего выброса вариационного ряда. С помощью критерия Граббса при проверке вариационного ряда не было выявлено выбросов ни справа, ни слева. Дальнейшее рассмотрение данных с использованием правила трех сигм также не выявило выбросов. Последним статистическим критерием, который использовался для выявления ошибок, был критерий Шовене. При проверке данных с его помощью была выявлена одна ошибка справа, выбросов слева обнаружено не было. По обобщенным результатам, а именно голосованием по большинству, трех статистических критериев можно сделать вывод, что вариационный ряд принадлежит одной генеральной совокупности. Представленные в статье критерии можно применять для анализа любых данных и на их основе делать соответствующие выводы, если в выборке были обнаружены ошибки.

Ключевые слова: статистический критерий, критерий Граббса, правило трех сигм, критерий Шовене, методы математической статистики, выброс.

THE USE OF STATISTICAL CRITERIA FOR EVALUATION TEST OF DATA ON PROPERTIES OF INORGANIC SUBSTANCES

I.D. Tarasenko[@], V.A. Dudarev

¹Moscow Technological University (Institute of Fine Chemical Technologies),
Moscow 119571, Russia

[@]Corresponding author e-mail: berkut257@gmail.com

Three statistical criteria were compared on the basis of the sampling of chemical data in the article. When the criteria were compared, the level of significance was taken as equal to 0.05, because this value is most often used in the technical calculations. Evaluation of data on the one-sided outliers of variational series was carried out by using statistical criteria. With the help of Grubbs criterion for checking variation series, outliers were found neither on the right nor on the left. Further examination of the data with the use of three-sigma rule also showed no outliers. The latest statistical criterion used to detect errors was Chauvenet's criterion. When checking the data with this criterion, one error on the right was detected. No outliers were found on the left. According to the generalized results, namely, the vote on the three most statistical criteria, it can be concluded that the variational series belongs to the same general totality. Criteria presented in the article can be applied for the analysis of any data and for making conclusions based on them, if errors were found in the sampling.

Keywords: statistical test, Grubbs criterion, three sigma rule, Chauvenet's criterion, methods of mathematical statistics, outlier.

Введение

Благодаря бурному развитию вычислительной техники и прикладной математики повсеместное применение методов машинного обучения и data mining для решения разнообразных задач прогнозирования в различных отраслях становится общепринятой тенденцией. Например, в неорганической химии успешно осуществляют не только прогнозирование возможности образования вещества, но и оценивают их свойства [1]. Однако при использовании этих методов анализа данных возникает много проблем, связанных как с качеством используемых данных, так и с особенностями применяемых методов.

В этой работе затрагивается проблема качества исходных данных, так как многие алгоритмы машинного обучения, например, широко используемый метод опорных векторов (SVM), являются чрезвычайно чувствительными даже к единичным выбросам в данных, используемых для обучения [2]. При наличии выбросов в данных эти алгоритмы становятся неэффективными, поэтому встает проблема предварительной проверки данных на предмет выбросов и их очистки перед использованием алгоритмов машинного обучения.

Методы математической статистики широко используются при анализе различных явлений и процессов. Если по результатам проведенных экспериментов требуется проверить некоторое предположение и сделать обоснованный вывод, то используется статистическая проверка гипотез.

Статистический критерий – это строгое математическое правило, по которому принимается или отвергается та или иная статистическая гипотеза на известном уровне значимости [3]. Для проверки гипотезы строится некоторая статистика. Если значение статистики попадает в область критических значений, то принимается альтернативная гипотеза, если не попадает – основная.

На практике важную роль играют статистические критерии, предназначенные для выявления аномальных результатов измерений (выбросов). Бывает так, что результат одного измерения поразительно расходится со всеми остальными. Когда такое происходит, нужно решить, является ли такой аномальный результат измерения следствием грубой ошибки и поэтому должен быть отброшен, или же это подходящий результат, который должен быть рассмотрен наряду с другими [4].

Результаты измерений, содержащие грубые ошибки, легко заметны и могут быть отделены без применения статистических методов. Применение статистических методов для выявления грубых ошибок целесообразно использовать лишь в спорных случаях, когда данные о качестве измерений недостаточно надежны.

Подавляющее большинство существующих критериев выявления выбросов в непрерывных статистических данных опирается на предположение о принадлежности наблюдаемых случайных величин нормальному закону распределения.

Экспериментальная часть

Статистические критерии

При решении задач статистического анализа и, в частности, при вычислении оценок параметров распределений проблема наличия в выборке аномальных измерений имеет важное значение. Присутствие единственного аномального наблюдения может приводить к оценкам, которые совершенно не согласуются с выборочными данными.

При сравнении критериев приняли уровень значимости равным 0.05, поскольку это значение чаще всего используется в технических измерениях. Для статистического анализа были взяты химические данные неорганических веществ по некоторым признакам (рис. 1). Оценивали на грубую ошибку только один односторонний выброс вариационного ряда.

1. Критерий Граббса

Одним из статистических критериев, позволяющих выявить аномальные измерения, является критерий Граббса.

Была получена выборка (около 700 значений) по значениям химического потенциала (Е8-1), и по ней построен вариационный ряд.

Проверяемая гипотеза H_0 заключается в том, что все $x_1, x_2, x_3, \dots, x_n$ принадлежат одной генеральной совокупности. При проверке на выброс наибольшего значения выборки альтернативная гипотеза H_1 заключается в том, что $x_1, x_2, x_3, \dots, x_{(n-1)}$ принадлежат одному закону, а x_n – некоторому другому, существенно сдвинутому вправо. При проверке на выброс наибольшего значения выборки статистика критерия Граббса имеет вид: [5]

$$G_{\max} = \frac{(x_n - \bar{x})}{S} \quad (1.1)$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.2)$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (1.3)$$

При проверке на выброс наименьшего выборочного значения конкурирующая гипотеза H_1 предполагает, что x_1 принадлежит некоторому другому закону, существенно сдвинутому влево. В данном случае вычисляемая статистика принимает вид:

Состав	E8-1	I8-2	S6-3	S5-4	E2-5	E5-6	E6-7	E7-8
CaYAlO ₄	3,2	17,2	1,22	2,42	1,2	615,6	1181	1980
CaLaAlO ₄	3,17	13,5	1,34	2,42	1	538,1	1067	1851
CaCeAlO ₄	3,18	11,4	1,23	2,4	1,05	527,8	1047	1940
CaPrAlO ₄	3,19	12,5	1,2	2,39	1,05	523	1018	2090
CaNdAlO ₄	3,19	16,5	1,17	2,37	1,05	529,7	1034	2128
CaSmAlO ₄	3,2	13,3	1,11	2,33	1,05	543,2	1068	2285
CaEuAlO ₄	3,2	13,9	1,1	2,29	1	547,1	1086	2425
CaGdAlO ₄	3,2	10,6	1,09	2,22	1,1	592,4	1186	1999
CaTbAlO ₄	3,21	11,1	1,08	2,2	1,1	564,4	1112	2122
CaDyAlO ₄	3,21	10,7	1,07	2,18	1,1	572,2	1126	2230
CaHoAlO ₄	3,22	16,2	1,04	2,19	1,1	580,8	1139	2221
CaErAlO ₄	3,22	14,3	1,02	2,15	1,1	588,6	1151	2207
CaTmAlO ₄	3,22	16,8	1,01	2,03	1,1	596,3	1163	2305
CaYbAlO ₄	3,22	34,9	1	1,99	1,05	603,4	1174	2408
CaLuAlO ₄	3,22	16,4	0,99	1,94	1,15	523,5	1341	2054
SrYAlO ₄	3,2	17,2	1,22	2,42	1,2	615,6	1181	1980
SrLaAlO ₄	3,17	13,5	1,34	2,42	1	538,1	1067	1851
SrCeAlO ₄	3,18	11,4	1,23	2,4	1,05	527,8	1047	1940

Рис. 1. Таблица химических данных по признаку.

$$G_{\min} = \frac{(\bar{x} - x_1)}{S} \quad (1.4)$$

Максимальные и минимальные значения считаются выбросом, если $G_{\max} \gg G_{\text{крит.}}$ или $G_{\min} \gg G_{\text{крит.}}$, при заданном уровне значимости.

Проверка выборки с значениями химического потенциала (E8-1) критерием Граббса не выявила промахов в измерениях (рис. 2).

Для поиска промахов редко используют только один критерий. Зачастую выборку экспериментальных данных проверяют по нескольким критериям.

2. Правило трех сигм

Практически все значения нормально распределенной случайной величины лежат в интервале

$-3\sigma < \bar{x} < 3\sigma$. Более строго – приблизительно с вероятностью 0.9973 значение нормально распределенной случайной величины лежит в указанном интервале (при условии, что величина истинная, а не полученная в результате обработки выборки) [6]:

$$P = (-\sigma < \bar{x} < \sigma) = 0,683 \quad (2.1)$$

$$P = (-2\sigma < \bar{x} < 2\sigma) = 0,954 \quad (2.2)$$

$$P = (-3\sigma < \bar{x} < 3\sigma) = 0,997 \quad (2.3)$$

Почти достоверно, что случайная величина (ошибка) не отклонится от математического ожидания по абсолютной величине больше, чем на 0.003. Это предположение называют правилом трех сигм [7]:

Вещество	Признак	Критерий Граббса	
AgMn ₂ O ₄	4,45		
BaAl ₂ O ₄	4,2	P=	0,95
BaAm ₂ O ₄	3,8	a=	0,05
BaAu ₂ O ₄	5,15	n=	775
BaB ₂ O ₄	5,3	x _{ср} =	3,6495
BaBi ₂ O ₄	4,15	S ² =	0,4605
BaCe ₂ O ₄	3,18		
BaCeAlO ₄	3,18	max	min
BaCeCrO ₄	3,18	G=	2,9478 0,7067
BaCeFeO ₄	3,18		Не ошибка Не ошибка
BaCeMnO ₄	3,18		
BaCr ₂ O ₄	4,65	G _{крит} =	3,78
BaDy ₂ O ₄	3,21		
BaDyAlO ₄	3,21		
BaDyCoO ₄	3,21		
BaDyCrO ₄	3,21		

Рис. 2. Проверка выборки критерием Граббса.

$$3\sigma < |x_n - \bar{x}| \quad (2.4) \quad K = \frac{(x_i - \bar{x})}{S} \quad (3.1)$$

Данные (рис. 3) также были проверены с помощью правила трех сигм. Максимальное и минимальное значения оказались значимыми, т.е. не попали в зону выброса.

3. Критерий Шовене

Существует много критериев выявления и отбрасывания промахов, но ни один из них не является универсальным. Выбор критерия зависит от цели измерений, но решение отбросить какие-то данные, в конечном счете, всегда субъективно [7].

Согласно критерию Шовене, элемент выборки x_i объема n является выбросом, если вероятность его отклонения от среднего значения не больше $1/12n$.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.2)$$

$$S = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.3)$$

Если $K > K^*$, (K^* – критическое значение из таблицы Шовене), то значение x_i должно быть признано ошибкой (выбросом).

После проверки представленной выборки с помощью критерия Шовене был обнаружен один выброс справа (максимальное значение) (рис. 4).

Вещество	Признак	Правило 3 сигм	
AgMn2O4	4,45		
BaAl2O4	4,2	n=	775
BaAm2O4	3,8	$x_{cp}=$	3,6495
BaAu2O4	5,15	$s^2=$	0,4605
BaB2O4	5,3	$s=$	0,6786
BaBi2O4	4,15		
BaCe2O4	3,18		
BaCeAlO4	3,18	$ x_n - x_{cp} =$	max min
BaCeCrO4	3,18		2,0005 0,4795
BaCeFeO4	3,18		Не ошибка Не ошибка
BaCeMnO4	3,18	$3\sigma=$	2,0358
BaCr2O4	4,65		
BaDy2O4	3,21		
BaDyAlO4	3,21		
BaDyCoO4	3,21		
BaDyCrO4	3,21		

Рис. 3. Проверка выборки правилом трех сигм.

Вещество	Признак	Критерий Шовене	
AgMn2O4	4,45		
BaAl2O4	4,2	P=	0,95
BaAm2O4	3,8	a=	0,05
BaAu2O4	5,15	n=	775
BaB2O4	5,3	$x_{cp}=$	3,6495
BaBi2O4	4,15	$s=$	0,4605
BaCe2O4	3,18		
BaCeAlO4	3,18		
BaCeCrO4	3,18		
BaCeFeO4	3,18		
BaCeMnO4	3,18		
BaCr2O4	4,65		
BaDy2O4	3,21		
BaDyAlO4	3,21		
BaDyCoO4	3,21		
BaDyCrO4	3,21		

Рис. 4. Проверка выборки критерием Шовене.

Заключение

Для проверки данных по свойствам неорганических веществ были применены критерии Граббса, Шовене и правило трех сигм для обнаружения одного выброса в генеральной совокупности. На основании проведенных расчетов показано, что критерии могут давать разную оценку данным из представленной генеральной совокупности: по критерию Граббса и правилу трех сигм не оказалось выбросов ни справа, ни слева, а по критерию Шовене был выявлен один выброс справа. В связи с отличающимися результатами, полученными разными методами, необходим

механизм их обобщения для принятия окончательного решения относительно наличия выброса. При голосовании по большинству трех статистических критериев можно сделать вывод, что вариационный ряд принадлежит одной генеральной совокупности.

Представленные в работе критерии и подходы можно применять для анализа данных любой природы и на их основе делать выводы относительно наличия выбросов в выборке. Однако для принятия более обоснованного решения требуется увеличение количества используемых критериев и использование более продвинутых коллективных методов обобщения полученных результатов.

Список литературы:

1. Дударев В.А., Филоретова О.А., Брыкина Г.В. Методы распознавания образов в компьютерном конструировании неорганических соединений // Прикладная информатика. 2014. № 2 (50). С. 82–88.
2. Журавлев Ю.И., Рязанов В.В., Сенько О.В. «РАСПОЗНАВАНИЕ». Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2006. 176 с.
3. Тейлор Дж. Введение в теорию ошибок: пер. с англ. М.: Мир, 2009. 272 с.
4. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. М.: ФИЗМАТЛИТ, 2011. 816 с.
5. Боровков А.А. Теория вероятностей. М.: Эдиториал УРСС, 1999. 472 с.
6. Боровиков В. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов: 2-е изд. СПб: Питер, 2003. 688 с.
7. Буховец А.Г., Москалев П.В., Богатова В.П., Бирючинская Т.Я. Статистический анализ данных в системе R. Воронеж: ВГАУ, 2014. 124 с.

References:

1. Dudarev V.A., Filoretova O.A., Brykina G.V. Pattern recognition methods in computer construction of inorganic compounds // Prikladnaya informatika (Applied Informatics). 2014. № 2 (50). P. 82–88. (in Russ.).
2. Zhuravlev Yu.I., Ryazanov V.V., Sen'ko O.V. «RECOGNITION». Mathematical methods. Software system. Practical applications. Moscow: FAZIS, 2006. 176 p. (in Russ.).
3. Teylor Dzh. Introduction to the theory of errors: Transl. from English. Moscow: Mir, 2009. 272 p. (in Russ.).
4. Kobzar' A.I. Applied mathematical statistics. For engineers and scientists. Moscow: FIZMATLIT, 2011. 816 p. (in Russ.).
5. Borovkov A.A. Probability theory. Moscow: Editorial URSS, 1999. 472 p. (in Russ.).
6. Borovikov V. STATISTICA. Art on a computer analysis of the data: For professionals: 2nd ed. SPb: Piter, 2003. 688 p. (in Russ.).
7. Bukhovets A.G., Moskaev P.V., Bogatova V.P., Biryuchinskaya T.Ya. Statistical analysis of data in the system R. Voronezh: VGAU, 2014. 124 p. (in Russ.).

Об авторах:

Тарасенко Иван Дмитриевич, аспирант кафедры информационных систем в химической технологии Института тонких химических технологий ФГБОУ ВО «Московский технологический университет» (119571, Россия, Москва, пр-т Вернадского, д. 86).

Дударев Виктор Анатольевич, кандидат технических наук, доцент кафедры информационных систем в химической технологии Института тонких химических технологий ФГБОУ ВО «Московский технологический университет» (119571, Россия, Москва, пр-т Вернадского, д. 86).