Стоит отметить, что Office Open XML является форматом по умолчанию для приложений Microsoft Office 2007 и более поздних [5]. Однако, часто при документообороте встречаются документы с расширением DOC, созданные в пакете Microsoft Office 2003. Третий и четвертый методы решения не поддерживают работу с такими документами. В этом случае можно использовать конвертер, который первоначально преобразует файл с расширением DOC в файл формата Office Open XML с расширением DOCX. В качестве конвертера может использоваться пакет OpenOffice, обращение к которому производится из командной строки.

Список использованных источников

1. Миронов В.В. Информационная технология персонализации электронных документов Microsoft Office в WEB-среде на основе XML / В.В. Миронов, Г.Р. Шакирова, В.Э. Яфаев // Вестник Уфимского государственного авиационного технического университета. 2008. №2. С. 112-122.
2. Сухов К. Возможности языка PHP. Работа с приложениями посредством технологии COM / К. Сухов. Системный администратор. 2010. № 3. С. 70-74.
3. DOTNET Manual [Электронный ресурс] // The PHP Group: [web-сайт]. – Режим доступа: http://php.net/dotnet (дата обращения 12.04.2017).
4. PHPWord [Электронный ресурс] // CodePlex: хостинг проектов для открытого программного обеспечения: [web-сайт]. – Режим доступа: https://phpword.codeplex.com/ (дата обращения 12.04.2017).
5. Office Open XML [Электронный ресурс] // Википедия. Свободная энцеклопедия: [web-сайт]. – Режим доступа: https://ru.wikipedia.org/wiki/Office_Open_XML (дата обращения 12.04.2017).

УДК 004.91

С. А. Дианов, В. Г. Лисиенко
ФГАОУ ВО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина», г. Екатеринбург, Россия

ПОИСК ВЫБРОСОВ В МНОГОМЕРНОМ МАССИВЕ ДАННЫХ

Аннотация

В докладе изложена проблема наличия выбросов в базах данных производственных процессов. Рассмотрена процедура обработки данных, направленная на поиск выбросов. Произведено сравнение различных методов поиска выбросов.

Ключевые слова: нахождение выбросов, предварительная обработка данных, идентификация процесса, робастность.

Abstract

The problem of outliers in industrial processes databases is outlined. The outlier detection data preprocessing procedure is examined. The different ways of outlier detection are compared.

Keywords: outlier detection, data preprocessing, process identification, robustness.

Моделирование технологических процессов методом «чёрного ящика» требует сбора большого количества данных, полученных путём сохранения мгновенных значений факторов и откликов в базу данных. Такие методы моделирования подвержены значительному искажениям, связанным с содержащимися в выборке резко выделяющихся наблюдений. Они могут возникать из-за несовершенства измерительного оборудования, нестабильности процесса, человеческого фактора. Все эти обстоятельства ведут к затруднению идентификации производственного процесса. Поэтому перед моделированием эти данные

нужно подвергнуть подготовке, направленной на обнаружение и изъятие из выборки резко выделяющихся наблюдений.

Для их нахождения предложено большое количество методов, основанных на анализе разброса данных и квантили распределения. Но большинство из них, например, критерий Смирнова или критерий Диксона [1], предназначены для анализа каждой переменной промышленного процесса в отдельности, что может быть недостаточно эффективно, когда эти переменные тесно связаны, а кроме того, они не дают хороших результатов, когда выборка достаточно большая и выбросов в ней больше одного. Значение каждой переменной может и не выделяться из общего массива, но вкупе они могут искажать данные достаточно сильно, чтобы затруднить идентификацию. В этом случае возможно применение методов, основанных на анализе многомерного массива данных. Таких методов тоже достаточно много, но их объединяет одно свойство: каждая величина в рассматриваемом векторе должна быть нормально распределена. В противном случае эти методы могут не дать хороших результатов.

К разрабатываемому методу предъявляются следующие требования:

– простота математического аппарата и программной реализации;
– возможность обрабатывать данные в реальном времени.

Метод позволяет определить, принадлежит ли та или иная точка к данной выборке. Он основан на простом анализе среднего арифметического и стандартного отклонения. В случае, если рассматриваемая точка не принадлежит выборке, то её удаление значительно изменит стандартное отклонение. Пороговый уровень изменения стандартного отклонения (α) является настраиваемым параметром метода.

Имеется матрица X, состоящая из n строк (векторов) и m столбцов (переменных процесса). Для каждого столбца матрицы вычисляется среднее арифметическое и среднеквадратичное отклонение. Затем исходная матрица X масштабируется по формуле (1):

$$X'_{i,j} = \frac{(X_{i,j} - m_i)}{s_i}, \quad (1)$$

где m_i – среднее значение i-й переменной; s_i – стандартное отклонение i-й переменной.

Получается матрица, приведённая к единичному стандартному отклонению и нулевому среднему арифметическому. Это необходимо для того, чтобы влияние каждой переменной на результат было одинаковым. После этого находят Евклидову метрику каждого вектора по формуле (2):

$$L = \sqrt{\sum X_i'^2}. \quad (2)$$

За точку, подозреваемую в выбросе, принимают ту, для которой значение L будет наибольшим. Удаляют эту точку из матрицы X' и вычисляют для неё стандартное отклонение. Если в получившемся векторе стандартного отклонения содержится хотя бы одно число, меньшее α, то точку признают выбросом и удаляют её из матрицы X. После этого заново вычисляют матрицу X' и находят Евклидову метрику каждого вектора. Находят наибольшее значение в матрице L и исследуют соответствующую точку на выброс. Процесс продолжают до тех пор, пока в векторе стандартного отклонения, получившемся после удаления точки из матрицы X', не останется значений, меньших α.

Проверим этот метод на некоторой выборке, состоящей из девяти переменных и 600 наблюдений, $\alpha=0,97$. Что представляют из себя эти переменные — в данном случае не имеет значения. Фрагмент выборки представлен в таблице 1.

Каждая анализируемая величина подчиняется нормальному закону распределения, поэтому к ним можно применять представленный алгоритм. После выполнения алгоритма получен результат, что наблюдения под номерами [210, 211, 212, 380, 383, 387, 389, 393, 400] являются выбросами. Удалим эти наблюдения из выборки и сравним статистические характеристики исходной и полученной выборки.

Увеличение стандартного отклонения в некоторых переменных говорит о том, что в удалённых наблюдениях их значение было близко к среднему.

Таблица 1

Фрагмент исходной выборки

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
5950	68,9	620,2	142,2	19	1481	8,2	798,5	8290
5950	68,3	628,1	145,2	23,5	1501	8,8	795,3	8659
5950	67	627,3	139,3	22,5	1475	9,3	801,8	8826
5950	66,4	629,3	141	21,5	1471	8,7	795,3	8725
5950	64	626	140,6	20	1495	7,6	792	8663

Таблица 2

Статистические характеристики результата

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
Среднее (до)	6014,3	75,05	624,14	138,88	20,97	1489,93	9,64	794,54	9126,6
Среднее (после)	6016,1	75,1	625,4	138,96	20,96	1489,97	9,63	795,7	9128,0
Станд. откл. (до)	71,25	5,44	16,13	12,64	1,54	17,88	0,88	16,85	302,3
Станд. откл. (после)	70,37	5,42	11,88	12,67	1,55	17,92	0,87	13,94	301,5
Мин. (до)	5750	62,7	523,6	103,3	16	1403	7,3	711,7	7565
Мин. (после)	5750	62,7	546,2	103,3	16	1403	7,3	721	7565
Макс. (до)	6100	90,7	642,1	165,9	25,5	1546	11,9	836,5	9920
Макс. (после)	6100	90,7	642,1	165,9	25,5	1546	11,9	836,5	9920

По данным таблицы 2 можно сделать вывод, что в представленной выборке содержалось небольшое количество выбросов, которые слабо влияли на статистические характеристики. Сравним получившиеся результаты с результатами, полученными по методам, предложенным Zeng, Gao [2]. Метод Resampling Half Means (RHM) на этой выборке признал выбросами следующие точки: [211 212 380 387 389 393 400]. Все эти точки присутствуют в результатах, полученных на методе отброса по отклонению. Анализ статистических характеристик исходной и полученной выборки дал те же результаты. Однако этот метод стохастичен, так как в нём нужно несколько раз случайным образом извлекать половину исходной выборки. Поэтому он время от времени может давать разные результаты. Метод, предложенный теми же авторами, основанный на метрике Махаланобиса, признал выбросами следующие точки: [211 212 387 392 393 400 401 412], что так же согласуется с другими методами.

Таким образом, предложенный метод способен решать поставленную задачу, прост в программной реализации, его результаты согласуются с результатами, полученными с помощью других методов.

Список использованных источников

1. Методы планирования и обработки результатов инженерного эксперимента: учебное пособие / Н.А. Спирин, В.В. Лавров, Л.А. Зайнуллин, А.Р. Бондин, А.А. Бурыкин; под общ. ред. Н.А. Спирина. – Екатеринбург: ООО «УИНЦ», 2015. – 284 с.

2. Zeng J., Gao C. Improvement of identification of blast furnace ironmaking process by outlier detection and missing value imputation // Journal of Process Control. 2009. № 9 (19). С. 1519–1528.