УДК 004.052.42

Методика выбора параметров и интерпретации результатов анализа выбросов в данных систем поддержки принятия решений

© В.И. Кузовлев, А.О. Орлов

МГТУ им. Н.Э. Баумана, Москва, 105005, Россия

Описана модель анализа категориальных атрибутов данных. Модель построена на вычислении показателя локальной аномальности LOF, расчете расстояний между значениями категориальных атрибутов с использованием формулы инверсной гравитации, понятиях плотности объектов и ядра. Обнаружена зависимость результатов работы модели от параметра k, характеризующего число ближайших объектов при расчете показателя LOF. Предложены интервалы значений параметра k, показан вариант применения этих интервалов при определении лингвистических переменных для использования в создании правил нечеткого вывода с целью обеспечения гибкости при выборе параметра k и возможности нечеткой интерпретации значений показателя LOF.

Ключевые слова: *показатель локальной аномальности, LOF, выбросы в данных, аномалии в данных, категориальные атрибуты.*

Введение. В системах поддержки принятия решений (СППР) большое значение уделяется формированию суждений о будущих фактах (прогнозам) на основе анализа статистических данных. Такой анализ данных называется прогнозным. Объекты генеральной совокупности представляют собой экземпляры некоторых сущностей, обладающие одинаковым набором атрибутов. Значения этих атрибутов анализируются для выявления закономерностей среди всех объектов генеральной совокупности (далее — объекты данных). Выбросами, или аномалиями, называются такие объекты данных, которые не удовлетворяют качествам, характерным для большинства других объектов генеральной совокупности. Поскольку каждый объект данных обладает рядом атрибутов, можно говорить о степени схожести объектов, основываясь на сравнении всех значений соответствующих атрибутов этих объектов. Большинство методов поиска выбросов в данных построены на вычислении расстояний между объектами данных [1]. В [2, 3] описывается метод поиска выбросов, основанный на расчете показателя локальной аномальности *LOF* [4]. Описаны преимущества данного метода. Одним из важных преимуществ является возможность расчета степени аномальности каждого объекта данных. Это позволяет гибко оценивать результат анализа в отличие

от методов, однозначно определяющих принадлежность объектов к аномалиям.

При использовании метода поиска выбросов возникают две проблемы. Во-первых, поскольку метод дает числовую оценку степени аномальности объектов, необходимо вводить некоторые дополнительные критерии, идентифицирующие выбросы. В данном случае удобно использовать механизмы теории нечетких множеств. Этапом дефаззификации при этом может являться переход от значения степени аномальности объекта данных к принятию решения о принадлежности его к выбросам.

Во-вторых, поскольку метод *LOF* основан на широко известном методе «*k* ближайших соседей», возникает задача выбора параметра *k*. В [4] даются общие рекомендации по выбору параметра *k*, однако авторы предлагают делать выбор отдельно для каждой задачи с учетом специфики анализируемых данных, их количества, прогнозируемого количества возможных выбросов и т. д.

В статье предложена методика формирования правил нечеткого вывода на основе параметра *k* при использовании показателя локальной аномальности для идентификации выбросов в данных.

Модель анализа. Все объекты генеральной совокупности *G* имеют одинаковый набор атрибутов $A = \{A_1, \ ..., \ A_n\}$. Каждый атрибут имеет некоторое количество уникальных значений $D(A_i) = \{x_{i1}, x_{i2}, \ ..., \ x_{im}\}$. В построенной модели объектами анализа являются значения отдельно взятого атрибута. Вообще, атрибуты могут иметь числовые или категориальные значения. В данной статье рассматриваются категориальные атрибуты. Они представляют больший интерес по сравнению с числовыми, поскольку заранее неизвестна их принадлежность каким-либо шкалам. Поэтому они требуют дополнительных процедур расчета расстояний между собой.

Для расчета расстояний между значениями категориального атрибута использовалась формула, предложенная в [4]:

$$\mathrm{dist}_{A_n}(x_i, x_j) = \sqrt{\frac{f_n(x_i) + f_n(x_j)}{f_n(x_i) f_n(x_j)}}, \qquad (1)$$

где A_n — категориальный атрибут, принимающий значения $D(A_n) = \{x_1, ..., x_p\}$; $f_n(x)$ — количество объектов генеральной совокупности, атрибут A_n которых принимает значение *x*.

Формула (1) называется формулой инверсной гравитации. Если представить объекты анализа как шарообразные тела, то частота $f_n(x)$ появления значения *x* атрибута A_n среди объектов генераль-

ной совокупности является массой. Введем параметр ρ, характеризующий плотность объектов. Будем считать плотность всех объектов одинаковой. Тогда, изменяя ρ, можно регулировать объем тел и, соответственно, площадь их проекций.

Если пересечение объектов x_i, x_j в некотором пространстве $W: x_i \cap x_j \neq \varnothing$, тогда $x_i \in C$ и $x_j \in C$. Множество C всех объектов, имеющих пересечения, называется ядром в пространстве W:

$$C = \left\{ x_1, x_2, ..., x_k \middle| \left(\bigcup_{i,j=1}^{k} \left(x_i \cap x_j \right) \right) \neq \varnothing \right\}. \quad (2)$$

Если представить множество C на плоскости, то $S(C)$ — есть площадь фигуры C. Тогда очевидны следующие выражения:

$$S(C) \leq \sum_{i=1}^{k} S(x_i),$$

где $x_i \in C$;

$$S(C) \leq \sum_{i=1}^{p} S(x_i) = S\left(D(A_n)\right),$$

где $p = \left|D(A_n)\right|$.

Работа построенной модели состоит из трех этапов. На первом этапе по формуле (1) рассчитываются расстояния между всеми объектами анализа. Так же вычисляются показатели локальной аномальности LOF для каждого объекта. На втором этапе происходит автоматический анализ среднего показателя $\overline{LOF}$ для объектов ядра, а также отношения площади фигуры ядра к общей площади фигур объектов S_{rel}:

$$\overline{LOF} = \frac{\sum_{i=1}^{|C|} LOF(x_i)}{|C|}; \quad (3)$$

$$S_{rel} = \frac{S(C)}{S\left(D(A_n)\right)}. \quad (4)$$

Параметр плотности объектов ρ уменьшается с заданным шагом, который автоматически корректируется по мере продвижения процесса анализа. При уменьшении плотности площадь объектов увеличивается, новые объекты попадают в пересечения, становясь частью

ядра. Снова рассчитывается средний показатель $\overline{LOF}$ по формуле (3) и отношение площадей по формуле (4). Плотность ρ уменьшается до тех пор, пока все объекты не попадут в ядро, т. е. станет справедливо равенство $S_{rel} = 1$.

На третьем этапе формируется график зависимости среднего показателя локальной аномальности объектов ядра от отношения площадей фигуры ядра к общей площади объектов $\overline{LOF}(S_{rel})$. Вся процедура повторяется для разных значений параметра k, характеризующего количество ближайших объектов при расчете показателя LOF.

Визуальное представление модели реализовано средствами библиотеки D3js [6] и показано на рис. 1.

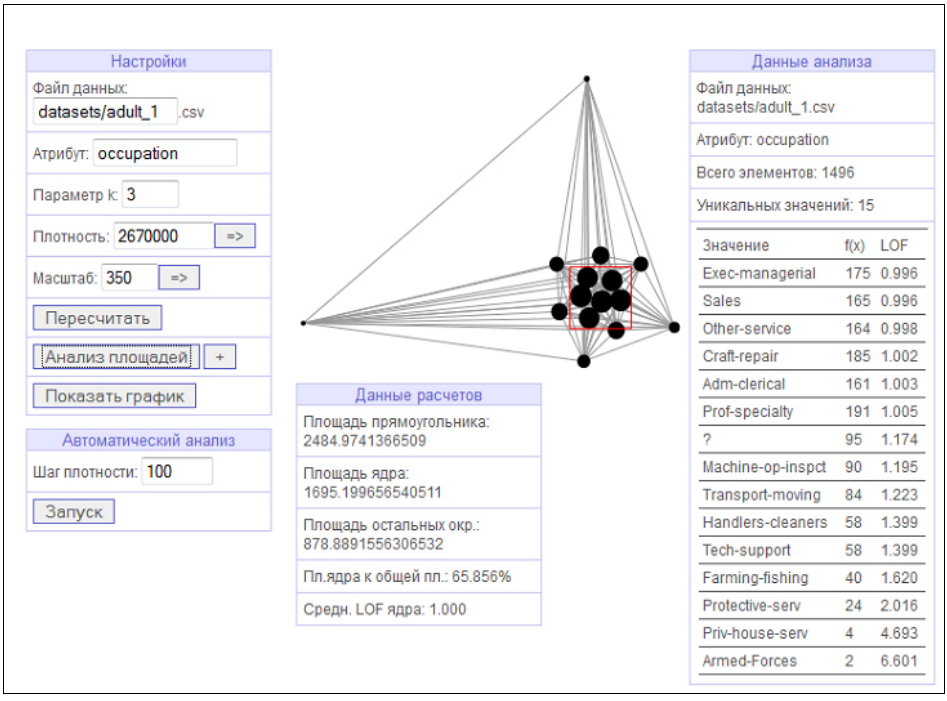


Рис. 1. Модель анализа значений категориальных атрибутов

Расчет площади фигуры, состоящей из пересечения набора окружностей, является весьма нетривиальной задачей [7]. Поэтому для расчета площади ядра использовался алгоритм Монте-Карло [8] с числом точек, равным 100 000.

Результаты моделирования. Для моделирования использовались наборы данных Калифорнийского университета: набор данных о флагах стран [9], содержащий 148 записей, а также часть набора данных о взрослом населении США [10], содержащая 1496 записей. Для

анализа были взяты некоторые категориальные атрибуты данных наборов, представленные в табл. 1.

Таблица 1

Данные анализа

Набор данных	Атрибут	Количество уникальных значений
Adults	Education (Образование)	16
Adults	Marital-status (Семейный статус)	7
Adults	Occupation (Сфера деятельности)	15
Adults	Native-country (Родная страна)	32
Flags	Mainhue (Превалирующий цвет)	8

Для каждого атрибута из табл. 1 формировалась модель, проводился расчет показателя LOF всех значений атрибута, строился график изменения отношения $\overline{LOF}$ точек ядра к относительной площади фигуры ядра (рис. 2). Анализ проводился для разных значений параметра k. На каждом графике определялась некоторая точка X, в которой начинался рост функции. Определялся $\overline{LOF}$ ядра в точке X, а также разброс ΔLOF точек ядра. Точки, не вошедшие в ядро в точке X, считаются выбросами при данном k.

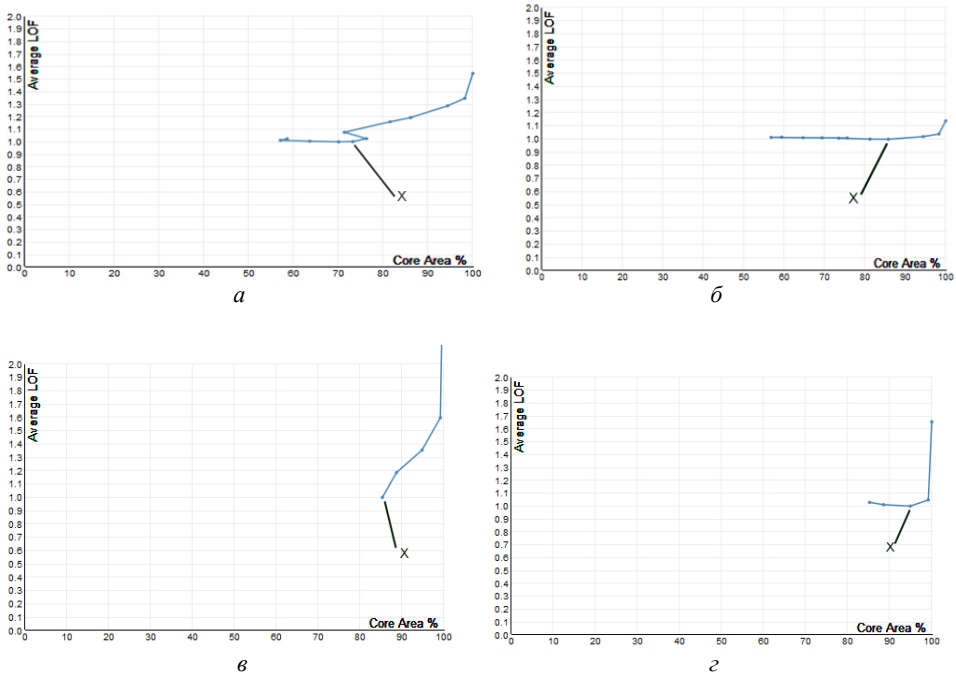
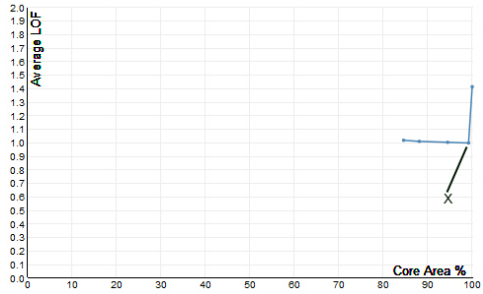
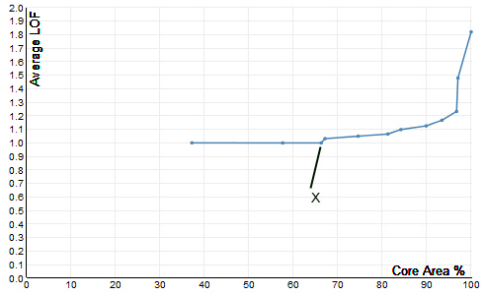


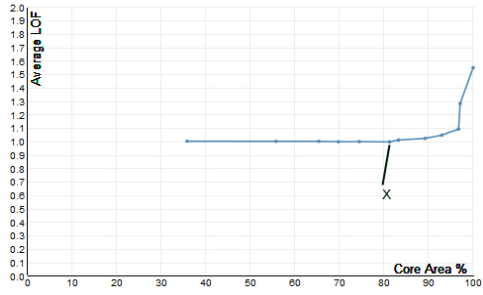
Рис. 2. Анализ $\overline{LOF}\left(S_{rel}\right)$ атрибутов



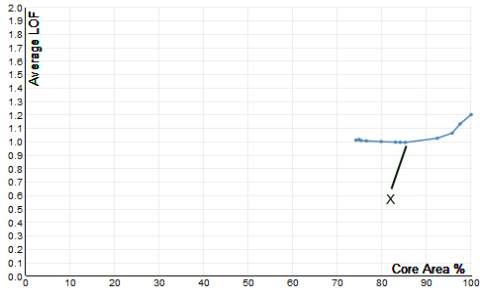
д



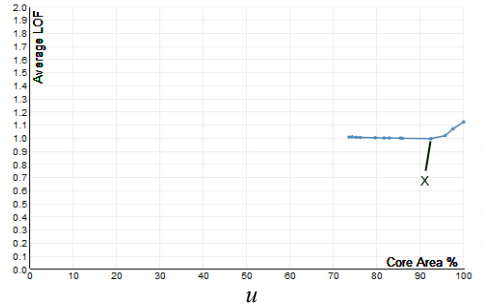
е



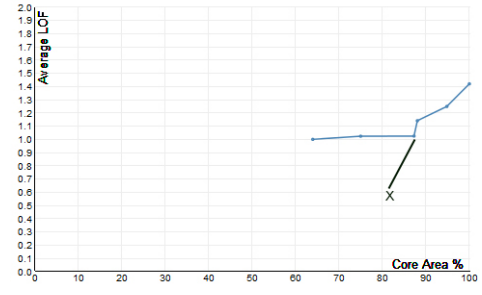
ж



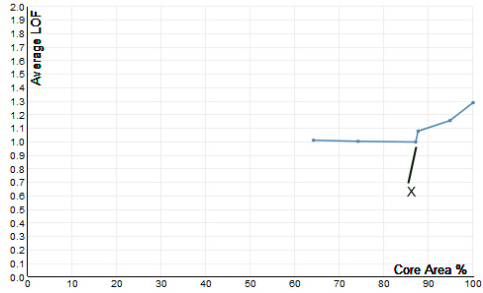
з



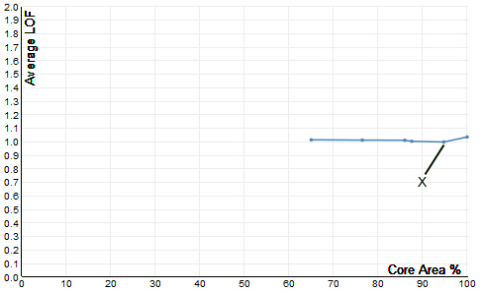
и



к



л



м

Рис. 2. Анализ $\overline{LOF}\left(S_{rel}\right)$ атрибутов (окончание)

В табл. 2 приведены результаты моделирования, отображенные на рис. 2.

Таблица 2

Результаты анализа атрибутов

Рисунок	Набор данных	Атрибут	k	$\overline{LOF}$ ядра	ΔLOF ядра	Количество выбросов
2, *а*	Adult	Education	5	1,003	0,056	9
2, *б*	Adult	Education	10	0,999	0,035	4
2, *в*	Adult	Marital-status	2	1,000	0,033	4
2, *г*	Adult	Marital-status	4	1,001	0,078	2
2, *д*	Adult	Marital-status	5	1,000	0,045	1
2, *е*	Adult	Occupation	3	1,000	0,009	9
2, *ж*	Adult	Occupation	7	1,001	0,024	5
2, *з*	Adult	Native-country	10	0,997	0,036	18
2, *и*	Adult	Native-country	16	0,998	0,022	11
2, *к*	Flags	Mainhue	2	1,026	0,120	3
2, *л*	Flags	Mainhue	4	1,000	0,053	3
2, *м*	Flags	Mainhue	6	1,000	0,052	1

Результаты моделирования подтверждают, что при увеличении параметра k график зависимости среднего показателя локальной аномальности объектов ядра от относительной площади фигуры ядра $\overline{LOF}\left(S_{rel}\right)$ становится более пологим. Это означает, что большее количество объектов попадает в ядро и меньше точек идентифицируются как выбросы. Таким образом, параметр k можно рассматривать как «регулятор» степени жесткости идентификации выбросов. Чем выше значение k, тем «мягче» анализ и меньше объектов будут отнесены к выбросам.

Пусть $p = \left|D\left(A_n\right)\right|$, количество уникальных значений атрибута A_n. При $k = p - 1$ график превратится в точку, поскольку в ядро попадут все объекты, т. е. $\left|C\right| = \left|D\left(A_n\right)\right|$.

Основываясь на данных результатов моделирования, можно сопоставить параметр k с количеством значений p, например, следующим образом. Введем лингвистическую переменную «анализ выбросов», характеризующуюся следующей пятеркой:

$$\left\{x = \text{"анализ выбросов"}, T\left(x\right), X, G = \varnothing, M\right\},$$

где $T\left(x\right) = \left\{x_1 = \text{"жесткий"}, x_2 = \text{"средний"}, x_3 = \text{"мягкий"}\right\}$ — множество значений переменной; $X = \left[1, p - 1\right]$ — интервал числовых значений.

Один из возможных способов определения функции принадлежности M изображен на рис. 3.

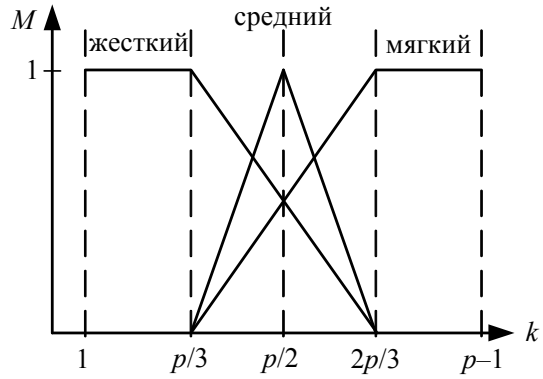


Рис. 3. Функция принадлежности лингвистической переменной «анализ выбросов»

По результатам проведенных исследований разработана методика выбора параметров для анализа данных систем поддержки принятия решений на предмет выбросов в категориальных атрибутах, а также последующей интерпретации результатов анализа. Разработанная методика состоит из следующих шагов:

Шаг 1. Формирование исходных данных для анализа. Файл представляет собой набор значений некоторого отдельно взятого категориального атрибута, являющийся подмножеством генеральной совокупности. При этом каждое значение записывается в новой строке, а первой строкой является название атрибута. Ясно, что количество строк в исходном файле соответствует мощности генеральной совокупности плюс один.

Шаг 2. При помощи модели, изображенной на рис. 1, проводится анализ значений категориального атрибута. При этом начальная плотность должна быть задана из тех соображений, чтобы в момент начала анализа не существовало пересечений объектов, т. е. ядро было пустым. Далее плотность будет автоматически регулироваться в процессе анализа.

Шаг 3. По результатам анализа данных будет построен график зависимости среднего *LOF* ядра от отношения площади ядра к суммарной площади всех объектов.

Шаг 4. Шаги 2–3 повторяются несколько раз для разных значений параметра k в диапазоне $[1, p-1]$, где p — количество уникальных значений рассматриваемого категориального атрибута. Таким образом, будет получен набор графиков зависимости среднего *LOF* ядра от его относительной площади.

Шаг 5. В зависимости от требуемой априори «жесткости» анализа выбрать значение параметра k, исходя из тех соображений, что чем

выше значение k, тем более «мягким» будет анализ, т. е. меньшее количество значений категориального атрибута будут идентифицированы как выбросы. Жесткость может быть выбрана на основе некоторых нечетких правил, использующих лингвистические переменные, как, например, на рис. 3.

Шаг 6. В графике, соответствующем выбранному значению параметра k, определить точку X начала роста функции, как на рис. 2.

Шаг 7. Выбросами считать точки, не вошедшие в ядро в точке X.

Заключение. Проведенное исследование показало, что точки, входящие в ядро, имеют разброс показателя *LOF* в пределах одной десятой. При расширении границ ядра в него начинают попадать точки, являющиеся выбросами. В этот момент отношение среднего показателя *LOF* ядра к его относительной площади начинает расти, что расценивается построенной моделью как сигнал о попадании в ядро потенциального выброса. Замечено, что при увеличении параметра k, определяющего количество ближайших соседей, анализируемых при расчете показателя *LOF*, график зависимости среднего *LOF* ядра от его относительной площади становится более пологим, сигнал о появлении выбросов появляется позже. Полученные выводы позволяют интерпретировать значения показателя *LOF*, а также гибко выбирать параметр k на основе субъективных ожиданий эксперта средствами нечеткой логики. На основе результатов моделирования предложен вариант разбиения параметра k на интервалы

$$\left[1, \frac{p}{3}\right], \left[\frac{p}{3}, \frac{2p}{3}\right], \left[\frac{2p}{3}, p-1\right],$$

где p — количество уникальных значений рассматриваемого категориального атрибута.

Разработана методика выбора параметров и интерпретации результатов анализа данных на предмет выбросов в категориальных атрибутах. Методика обладает достаточной гибкостью за счет возможности использования правил нечеткого вывода при выборе параметров анализа на основе нечетко заданной «жесткости» анализа.

Дальнейшие исследования могут быть посвящены развитию анализа категориальных атрибутов данных на основе построенной модели, формированию системы правил нечеткого вывода с использованием предложенного способа определения параметра k, а также разработке нечеткого алгоритма автоматической интерпретации показателя локальной аномальности.

ЛИТЕРАТУРА

[1] Chandola V., Banerjee A., Kumar V. Anomaly Detection: A Survey. *ACM Computing Surveys*, 2009, vol. 41, no. 3, article 15, p. 58.

[2] Кузовлев В.И., Орлов А.О. Метод выявления аномалий в исходных данных при построении прогнозной модели решающего дерева в системах поддержки принятия решений. *Наука и образование,* 2012, № 9. DOI: http://dx.doi.org/10.7463/0912.0483269

[3] Кузовлев В.И., Орлов А.О. Прогнозный анализ данных методом ID3O. *Наука и образование*, 2012, № 10. DOI: http://dx.doi.org/10.7463/1012.0483286

[4] Breunig M., Kriegel H.-P., T. Ng R., Sander J. LOF: Identifying Density-Based Local Outliers. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 93–104.

[5] Орлов А.О. Проблема поиска расстояний между значениями категориальных атрибутов при обнаружении выбросов в данных. *В мире научных открытий*, 2012, № 8.1, с. 142–155.

[6] *D3js — Data-Driven Documents JavaScript library*. URL: http://d3js.org (дата обращения 15.03.2013).

[7] Librino F., Levorato M., Zorzi M. An algorithmic solution for computing circle intersection areas and its applications to wireless communications: *In proceedings of Modeling and Optimization in Mobile*, *Ad Hoc, and Wireless Networks*, 2009, pp. 1–10.

[8] Метод Монте-Карло. *Википедия*. URL: http://ru.wikipedia.org/wiki/Метод_Монте-Карло (дата обращения 15.03.2013).

[9] Flags Data Set. *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml/datasets/Flags (дата обращения 11.03.2013).

[10] Adult Data Set. *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml/datasets/Adult (дата обращения 11.03.2013).

Статья поступила в редакцию 28.06.2013

Ссылку на эту статью просим оформлять следующим образом:

Кузовлев В.И., Орлов А.О. Методика выбора параметров и интерпретации результатов анализа выбросов в данных систем поддержки принятия решений. *Инженерный журнал: наука и инновации,* 2013, вып. 11.URL: http://engjournal.ru/catalog/it/hidden/1045.html

Кузовлев Вячеслав Иванович — канд. техн. наук, доцент кафедры «Системы обработки информации и управления» МГТУ им. Н.Э. Баумана.

Орлов Антон Олегович — инженер, выпускник кафедры «Системы обработки информации и управления» МГТУ им. Н.Э. Баумана. e-mail: forewar@gmail.com