

# СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ИНТЕРВАЛОВ ЦЕНЗУРИРОВАНИЯ ДАННЫХ В ЗАДАЧАХ АНАЛИЗА КРЕДИТОСПОСОБНОСТИ ОРГАНИЗАЦИЙ

**Р. Ф. Джафаров**

*Белорусский государственный университет, г. Минск;*

*DRF95@mail.ru; науч. рук. – В. И. Малюгин, канд. физ.-мат. наук, доц.*

В статье решается задача кластерного анализа многомерных данных финансовой отчетности организаций на классы, соответствующие различным уровням кредитоспособности. Для обнаружения аномальных наблюдений используются алгоритмы машинного обучения. Для учета аномальных наблюдений предлагается процедура цензурирования, которая позволяет сохранить всю выборку для последующего проведения кластерного анализа.

**Ключевые слова:** анализ кредитоспособности; аномальные наблюдения; цензурирование; машинное обучение.

## ВВЕДЕНИЕ

В [1] была предложена статистическая методика оценки классов кредитоспособности нефинансовых организаций на основе данных финансовой отчетности. Результатом применения данной методики является построение системы статистических кредитных рейтингов на микроуровне (рейтинги организаций) и макроуровне (средние отраслевые рейтинги и интегральный показатель кредитоспособности экономики в целом). Статистические кредитные рейтинги организаций получаются с помощью алгоритмов кластерного анализа. Классификационные признаки (финансовые балансовые коэффициенты) формируются с использованием преобразований цензурирования выборки и нормирования классификационных признаков, целью которых является учет аномальных наблюдений, приведение признаков к единому масштабу измерения и получение однообразной содержательной интерпретации для всех признаков.

Для цензурирования выборки в [1] используются экспертные оценки границ цензурирования для достаточно большого числа исходных признаков и различных отраслей экономики. Априорное задание экономически обоснованных границ цензурирования не позволяет, однако, учесть особенности реальных данных, в частности ярко выраженную асимметрию распределений вероятностей многомерных наблюдений и наличие аномальных наблюдений.

В данной статье предлагается процедура построения статистических оценок границ интервалов цензурирования на основе алгоритмов машинного обучения, которая позволяет учесть указанные особенности ре-

альных данных и получить в итоговом 14-мерном пространстве признаков 4 класса кредитоспособности организаций.

## ОПИСАНИЕ ДАННЫХ

В настоящем исследовании используются данные по 562 белорусским предприятиям шести видов экономической деятельности (отраслей). Источником информации является бухгалтерская отчетность субъектов хозяйствования (бухгалтерский баланс, отчет о прибылях и убытках), представленная за отчетные периоды: 3 месяца, 6 месяцев, 9 месяцев, 12 месяцев. Данные предоставлены за период 1 кв. 2013 – 2 кв. 2017 гг.

## АЛГОРИТМЫ АНАЛИЗА АНОМАЛЬНЫХ НАБЛЮДЕНИЙ

**Алгоритм Isolation Forest (IF)** представляет собой эффективный способ нахождения аномальных наблюдений в пространствах с высокой размерностью. Алгоритм «изолирует» наблюдения, выбирая случайным образом признак и некое его случайное значение между минимальным и максимальным. Рекурсивное разбиение может быть представлено в виде дерева, а число разбиений, требуемое для изоляции отдельного наблюдения равно длине пути от корня до конечного узла дерева. Эта длина пути, усредненная вдоль всего ансамбля деревьев, является мерой аномальности и решающей функцией. Случайное разбиение выдаёт заметно короткие длины путей для аномальных наблюдений. Таким образом, когда ансамбль деревьев совместно выдает сравнительно короткие длины путей для отдельных наблюдений, то они с высокой вероятностью являются аномалиями [2].

**Алгоритм Local Outlier Factor (LOF)** демонстрирует хорошие результаты в пространстве с умеренно большой размерностью. Этот алгоритм вычисляет оценки, которые отражают степень аномальности наблюдений. Эти оценки описывают отклонение локальной плотности в окрестности данного наблюдения и в окрестностях рядом лежащих точек. Идея заключается в том, чтобы найти наблюдения, которые имеют значительно более низкую плотность, чем у своих соседей [3].

Плотность множества точек в локальной окрестности некоторой точки  $x \in \mathcal{R}^N$  ( $N \geq 1$ ) определяется по формуле:

$$p_K(x) = 1 / \left( \frac{\sum_{k \in K} \rho(x, k)}{|K|} \right), \quad (1)$$

где  $K$  – множество ближайших соседей точки  $x$ , а  $\rho(x, k)$  – расстояние от точки  $x$  до соседа  $k$ . Функция (1) называется локальной плотностью.

Степень аномальности наблюдения в соответствии с LOF-алгоритмом (LOF-оценка) рассчитывается по формуле:

$$LOF(x) = \frac{\frac{1}{|K|} \sum_{k \in K} p(k)}{p(x)}. \quad (2)$$

Таким образом, обычное наблюдение имеет LOF-оценку вида (2), близкую к единице, в то время как для аномального наблюдения это значение значительно больше единицы.

**Алгоритм Elliptic Envelop (EE)** осуществляет поиск аномальных наблюдений, предполагая, что данные без выбросов имеют известное распределение вероятностей (например, нормальное). Основываясь на этом предположении, алгоритм пытается определить вид распределения данных и старается найти аномальные наблюдения, то есть те наблюдения, которые находятся достаточно далеко от множества концентрации (кластера) основной части всех наблюдений в соответствии с предполагаемым распределением. Для случайного вектора признаков  $x_i = (x_{ij}) \in \mathcal{R}^N$  ( $i = 1, \dots, n$ ,  $j = 1, \dots, N$ ), распределенного по нормальному закону, расстояние до центра распределения определяется по формуле Махаланобиса:

$$d(x_i) = \sqrt{(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)}, \quad i = 1, \dots, n, \quad (3)$$

где для случайного вектора  $x_i \in \mathcal{R}^N$ :  $\mu = \mathbf{E}\{x_i\} \in \mathcal{R}^N$  – вектор математического ожидания,  $\Sigma = \mathbf{E}\{(x_i - \mu)(x_i - \mu)^T\}$  – ковариационная матрица.

По случайной выборке  $\{x_i\}$  объема находится робастная оценка матрицы  $\Sigma$  с использованием наблюдений, попадающих в доверительную область (гиперэллипсоид) [3]. По результатам анализа аномальных наблюдений для каждого из  $N$  признаков определяются статистически обоснованные максимальные и минимальные пороговые значения (границы цензурирования)  $(K_{j, \min}, K_{j, \max})$  ( $i = 1, \dots, N$ ).

## ПРОЦЕДУРА ЦЕНЗУРИРОВАНИЯ И НОРМИРОВКИ

Целью цензурирования и нормировки данных является замена аномально больших и малых значений коэффициентов  $\{x_i\}$  на соответствующие границы цензурирования  $(K_{i, \min}, K_{i, \max})$  ( $i = 1, \dots, N$ ). Для приведе-

ния к единому масштабу измерения и удобной интерпретации используются преобразования признаков вида [1, 5]:

$$\tilde{x} = \frac{x - K_{\min}}{K_{\max} - K_{\min}} \text{ или } \tilde{x} = \frac{K_{\max} - x}{K_{\max} - K_{\min}}.$$

По результатам проведенного исследования ЕЕ-алгоритм показал наибольшую эффективность в выявлении аномальных наблюдений и определении границ цензурирования. На рисунке 1 представлена гистограмма исходного распределения коэффициента К4 (коэффициент обеспеченности собственными оборотными средствами) с ярко выраженной асимметрией и аномальными наблюдениями в области малых значений. Рисунок 2 иллюстрирует результат применения процедуры цензурирования и нормировки с использованием статистических оценок интервалов цензурирования, полученных с помощью ЕЕ-алгоритма: цензурированные наблюдения отнесены к нижней границе интервала цензурирования; гистограмма распределения основной массы наблюдений демонстрирует близость к симметричному распределению.

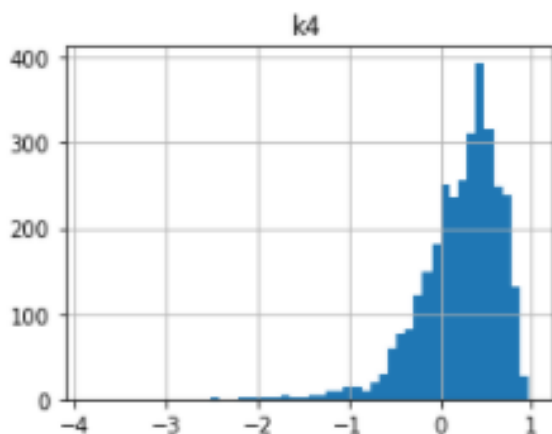


Рис. 1. Исходное распределение

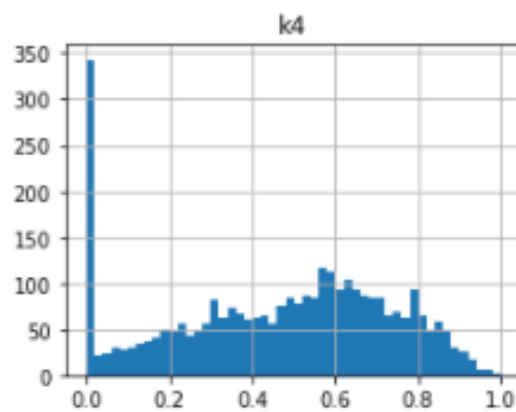


Рис. 2. Конечное распределение

### Библиографические ссылки

1. Система статистических кредитных рейтингов предприятий: методика построения, верификации и применения / В. И. Малюгин [и др.] // Банковский вестник. 2013. № 5. С. 4–73.
2. Liu F. T., Ming T. K., Zhou Z. H. Isolation Forest // Proceedings of the Eighth IEEE International Conference on Data Mining. 2009. P. 413–422.
3. Breunig M. M., Kriegel P. J., Ng R. T. Local Outlier Factor: Identifying Density-Based Local Outliers // Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, May 16–18, 2000. P. 93–104.
4. Rousseeuw P. J., Driessen K. V. A Fast Algorithm For The Minimum Covariance Determinant Estimator // Technometrics. 1999. № 41. P. 212–223.
5. Разработка модельного инструментария для оценки и прогнозирования кредитного риска банков на микроуровне : заключительный отчет о НИР / рук. В. И. Малюгин. Мн., 2018.