ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

ОБРАБОТКА ИЗОБРАЖЕНИЙ, РАСПОЗНАВАНИЕ РЕЧИ И ОБРАЗОВ

IMAGE PROCESSING, SPEECH AND PATTERN RECOGNITION

УДК 616.155.1-076.5:616.155.194.113-056.7

Поступила в редакцию 02.04.2019
Received 02.04.2019

Принята к публикации 20.05.2019
Accepted 20.05.2019

Статистическая классификация эритроцитов при наследственном сфероцитозе на основе спектральных признаков АСМ-изображений поверхностей клеток

И. Е. Стародубцев[1]✉, Ю. С. Харин[1, 2], М. С. Абрамович[1, 2]

[1]Белорусский государственный университет, Минск, Беларусь
✉E-mail: istarodubtsev.science@gmail.com
[2]Научно-исследовательский институт прикладных проблем математики и информатики Белорусского государственного университета, Минск, Беларусь

Аннотация. Предложен метод классификации эритроцитов по спектральным признакам изображений (микромасштабных карт физико-механических свойств), полученных сканированием поверхностей клеток на атомно-силовом микроскопе (АСМ). Для расчета признаков каждая линия сканирования исходного АСМ-изображения рассмотрена как реализация случайной последовательности и для нее применено дискретное преобразование Фурье. После сглаживания по полученной карте спектральных оценок построены информативные характеристики – медианы значений спектрограмм для каждой частоты. Проведена статистическая классификация эритроцитов двух типов (сфероцитов и дискоцитов) пациентов с наследственным сфероцитозом по полученным информативным характеристикам с помощью алгоритмов «деревья решений» и «бустинг на деревьях решений». Найден частотный интервал с наилучшей точностью классификации – более 82 % для алгоритма «бустинг на деревьях решений».

Ключевые слова: АСМ-изображения, эритроциты, карты физико-механических свойств, методы классификации, дискретное преобразование Фурье, деревья решений, бустинг на деревьях решений

Для цитирования. Стародубцев, И. Е. Статистическая классификация эритроцитов при наследственном сфероцитозе на основе спектральных признаков АСМ-изображений поверхностей клеток / И. Е. Стародубцев, Ю. С. Харин, М. С. Абрамович // Информатика. – 2019. – Т. 16, № 3. – С. 7–13.

Statistical classification of erythrocytes in hereditary spherocytosis based on the spectral features of cell surfaces' AFM images

Ivan E. Starodubtsev[1]✉, Yury S. Kharin[1, 2], Mikhail S. Abramovich[1, 2]

[1]Belarusian State University, Minsk, Belarus
✉E-mail: istarodubtsev.science@gmail.com
[2]Research Institute for Applied Problems of Mathematics and Informatics of the Belarusian State University, Minsk, Belarus

Abstract. The method of classification of erythrocytes (red blood cells) based on spectral features of the cell surface images (of physical-mechanical properties maps) obtained with an atomic-force microscope (AFM) is proposed. Each scan line of the original AFM image is considered as a random sequence realization and the

discrete Fourier transform is applied to compute its spectral features. The spectral estimates are smoothed on the map and the informative characteristics are computed as the medians of the spectrogram values for each frequency. The classification or two classes of erythrocytes (spherocytes and discocytes) taken from patients with hereditary spherocytosis was carried out by the obtained informative characteristics using the decision trees and boosted decision trees methods. The frequency interval was found with the best classification accuracy – over 82 % for the boosted decision trees method.

Keywords: AFM images, erythrocytes, maps of physical-mechanical properties, method of classification, discrete Fourier transform, decision trees, boosted decision trees

For citation. Starodubtsev I. E., Kharin Yu. S., Abramovich M. S. Statistical classification of erythrocytes in hereditary spherocytosis based on the spectral features of cell surfaces' AFM images. *Informatics*, 2019, vol. 16, no. 3, pp. 7–13 (in Russian).

Введение. Атомно-силовая микроскопия является современным методом медико-биологических исследований, позволяющим изучать на микро- и наноуровнях рельеф, структуру и физико-механические свойства поверхностей биологических клеток и, соответственно, определять их тип и состояние [1, 2].

Поверхность биологической клетки играет существенную роль в функционировании как самой клетки, так и организма в целом. Клетка через поверхность обменивается с окружающей ее средой веществом, энергией и информацией, взаимодействует с другими клетками. Поверхность биологической клетки имеет сложный рельеф, состав и структуру. Физико-механические свойства (упругие, адгезионные, фрикционные и др.) являются важными характеристиками поверхностей клеток. Клетки организма постоянно подвергаются действию механических сил (растяжению, сжатию, давлению и т. д.). Процессы, протекающие в биологических клетках, влияют на рельеф и физико-механические свойства их поверхностей, включая распределение (карты) физико-механических свойств. Эти изменения отражаются в АСМ-изображениях поверхностей клеток [3, 4].

Количественные характеристики изменений могут быть получены из АСМ-данных только после их соответствующей математической обработки, которая является необходимым этапом распознавания образов клеток. Распознавание образов включает поиск оптимальных (в определенном смысле) информативных признаков и построение оптимальных решающих правил для классификации исследуемых объектов, т. е. для отнесения их к определенным фиксированным классам [5]. Классами в контексте распознавания образов биологических клеток являются их морфологические типы и состояния, в том числе связанные с нормой и патологией клеток.

Современные подходы позволяют анализировать и количественно предсказывать функциональные свойства некоторых твердых поверхностей, основываясь на спектральных признаках их АСМ-изображений [6]. Разработка новых подходов, методов и алгоритмов распознавания поверхностей, позволяющих анализировать АСМ-изображения биологических клеток, является актуальной научно-исследовательской задачей.

Цель настоящей статьи заключается в статистической классификации эритроцитов различных типов (дискоцитов и сфероцитов) при наследственном сфероцитозе с помощью спектральных оценок АСМ-изображений карт физико-механических свойств микромасштабных участков поверхности.

АСМ-данные и их математическая модель. В работе анализировались АСМ-изображения карт физико-механических свойств, записанные в режиме сканирования torsion (карты сил трения скольжения), микромасштабных участков поверхностей эритроцитов, полученных от пациентов с наследственным сфероцитозом, т. е. патологических эритроцитов двух классов: дискоцитов и сфероцитов [7]. Использовались АСМ-изображения участков поверхностей сфероцитов (51 изображение) и дискоцитов (63 изображения).

Размер АСМ-изображений – $2{,}5 \times 2{,}5$ мкм, разрешение – 256×256 пикселов ($N = 256$). Изображения записаны на АСМ НТ-206 (производитель – ОДО «Микротестмашины», Беларусь) [2].

АСМ-изображение поверхности биологической клетки представляет собой двумерный массив $z = z(x, y)$, где x – координата по вертикали, $x \in \{1, 2, ..., N\}$; y – координата по горизонтали,

$y \in \{1, 2, ..., N\}$; z – значение силы трения в точке (x, y), который описывает карту локальных значений физико-механических свойств (сил трения скольжения). АСМ-изображение размером $N \times N$ точек можно рассматривать как совокупность из N одномерных массивов $z = z^{(y)}(x)$ по N точек в каждом, расположенных на расстоянии шага сканирования вдоль оси y (N – четное число).

АСМ-данные обрабатывались с помощью программного комплекса, разработанного авторами статьи на языке С++ в среде Borland C++ Builder 6.

Формирование информативных признаков. Каждый одномерный массив $z = z^{(y)}(x)$ при фиксированном y можно рассматривать как реализацию случайной последовательности $z = z_x^{(y)}, x = \{1, 2, \ldots, N\}$. Определим для этой последовательности дискретное преобразование Фурье:

$$X^{(y)}(\omega_k) = \sum_{n=1}^{N} (z_n^{(y)} - \overline{z}^{(y)})\, e^{-j\frac{2\pi kn}{N}}, \; k = 0, 1, ..., N-1, \tag{1}$$

где

$$\overline{z}^{(y)} = \frac{1}{N} \sum_{n=1}^{N} z_n^{(y)} \tag{2}$$

есть выборочное среднее по вертикали x при фиксированном y, $\omega_k = 2\pi \frac{k}{L}$ – частота, L – длина анализируемого интервала по оси x (сумма длин элементарных интервалов – шагов сканирования).

На основе выборочного спектра (1) построим периодограмму

$$r^{(y)}(\omega_k) = \left| X^{(y)}(\omega_k) \right|^2, \; k = 0, 1, ..., N-1. \tag{3}$$

Значения периодограммы $r^{(y)}(\omega_k)$ при $k = \frac{N}{2} + 1, ..., N$ исключим из рассмотрения, так как они повторяют значения периодограммы при $k = 0, ..., \frac{N}{2} - 1$.

Для получения оценки спектральной плотности $R^{(y)}(\omega_k)$ при фиксированном значении y полученная периодограмма $r^{(y)}(\omega_k)$ сглаживалась с помощью окна Даниэля размером m (m нечетное) с равными весами [8]:

$$R^{(y)}(\omega_k) = \frac{\sum_{l=k-\frac{m}{2}}^{k+\frac{m}{2}} R^{(y)}(\omega_l)}{m}, \; k = \frac{m}{2}, \frac{m}{2} + 1, ..., \frac{N}{2} - \frac{m}{2}. \tag{4}$$

Совокупность из N кривых $R^{(y)}(\omega_k)$, построенных согласно (4), при изменении координаты $y = \{1, 2, ..., N\}$ представляет собой карту, описывающую изменение спектральных оценок АСМ-изображения вдоль горизонтальной оси y. Для каждой частоты ω_k вычислялась медиана спектральной плотности по N ее значениям вдоль оси y:

$$\tilde{R}(\omega_k) = \mathrm{Med}\{R^{(1)}(\omega_k), ..., R^{(N)}(\omega_k)\}, \; k = 0, ..., \frac{N}{2}, \tag{5}$$

которую для краткости будем называть спектрограммой исследуемой клетки.

Полученный набор спектральных отсчетов $\{\tilde{R}(\omega_k): k = 0, 1, ..., \frac{N}{2}\}$ будем использовать как информативную характеристику исходного АСМ-изображения. При N = 256 для распознавания АСМ-изображений эритроцитов предлагается частотную область $\{\omega_k\}$ кривой $\tilde{R}(\omega_k)$ разбить на интервалы по l точек вида $\Omega_u = [\omega_{lu+1}, \omega_{l(u+1)}]$ и классификацию проводить по значениям признаков $\{\tilde{R}(\omega_k): \omega_k \in \Omega_u\}$, вычисленных только для частот заданного интервала. В данном случае, учитывая ограниченный объем выборки, удается избежать «эффекта переобучения». При этом частоты $\{\omega_k : k = 71, ..., 128\}$ исключаются из рассмотрения, так как соответствующие им периоды $T_k = \frac{2\pi}{\omega_k}$ меньше точности АСМ-прибора, на котором были получены АСМ-данные.

Алгоритмы классификации. Исходные АСМ-данные были нормализированы путем деления значений $z(x, y)$ на 10^3. Вместо исходных значений z для каждой линии сканирования рассматривались z'-разности между соседними значениями по оси x, деленные на значение среднеквадратичного отклонения для каждой из этих линий: $z'(x, y) = \frac{z(x+1, y) - z(x, y)}{\sqrt{D^{(y)}}}$, $x = \{1, 2, ..., N-1\}$. Затем с помощью библиотеки fftw (URL: http://www.fftw.org/) были вычислены Фурье-преобразования (1), (2) и спектральные признаки (3)–(5).

Для выделения интервалов частот $\{\Omega_u\}$, с использованием которых можно получить наибольшую точность классификации различных типов эритроцитов, первоначально применялся алгоритм классификации «деревья решений». Для построения деревьев решений использовался алгоритм C&RT [9], в котором каждая вершина дерева представляет собой правило классификации, связанное с одним признаком. Для выбора наилучшего из всех возможных вариантов ветвления, т. е. выбора признаков, по которым строилось дерево решений, использовался критерий Джини [9].

Выбор оптимального размера дерева классификации определялся с помощью кросс-проверки [10]. Этот вид проверки рекомендуется применять в том случае, когда нет отдельной экзаменационной выборки, а обучающая выборка имеет недостаточно большой объем.

Для повышения точности классификации в последнее время широко используются ансамбли алгоритмов классификации. Под ансамблем алгоритмов классификации понимается набор базовых классификаторов, результаты классификации которых затем объединяются и формируют решение агрегированного классификатора. Для построения ансамбля применялся бустинг на деревьях решений [9]. Идея бустинга состоит в том [9], что классификаторы ансамбля строятся последовательно и на каждой итерации происходит перевзвешивание (коррекция) наблюдений таким образом, чтобы соответствующий классификатор делал минимум ошибок на тех наблюдениях, на которых часто делали ошибки классификаторы, построенные на предыдущих итерациях алгоритма. Алгоритм градиентного бустинга на деревьях решений позволяет строить аддитивную функцию классификации в виде суммы деревьев решений итерационно по аналогии с методом градиентного спуска:

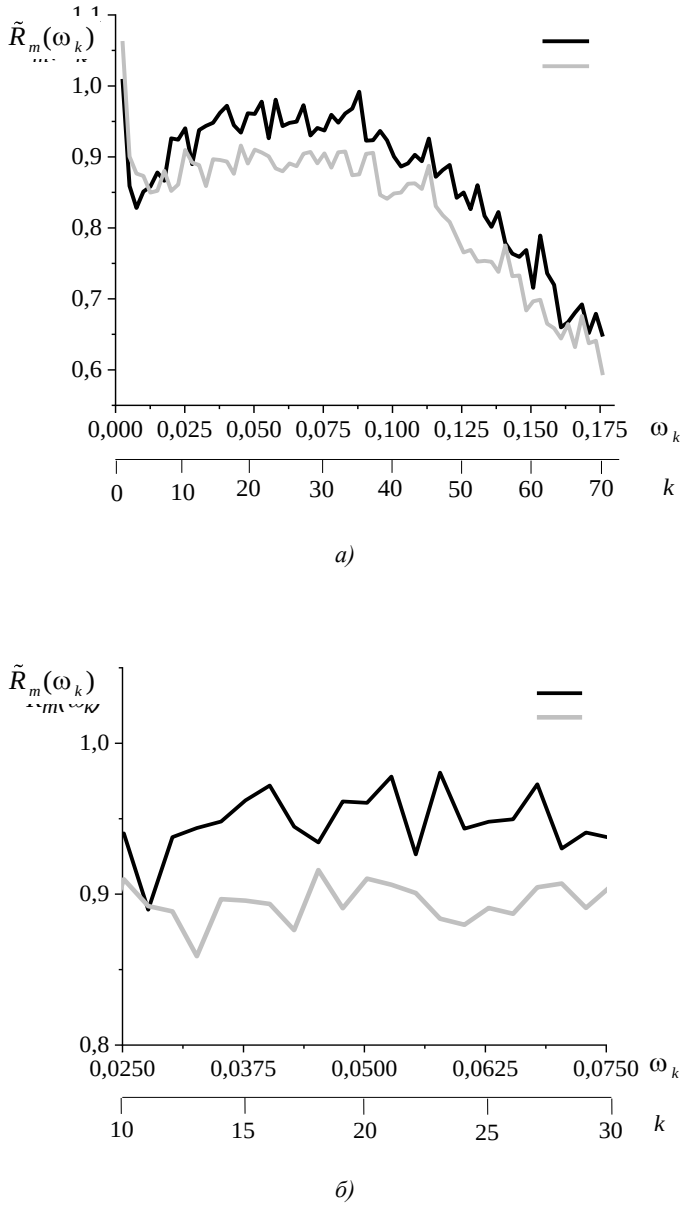
$$f(x) = h_0 + \upsilon \sum_{i=1}^{m} d_i(x), \tag{6}$$

где h_0 – константа; $\upsilon \in [0,1]$ – параметр, регулирующий скорость обучения и влияние отдельных деревьев на всю модель (коэффициент обучения); $d_i(x)$ – деревья решений; m – число деревьев решений. Ключевыми параметрами для бустинга на деревьях решений являются количество деревьев, максимальная глубина каждого дерева и коэффициент обучения.

Численные результаты. В связи с тем что объем обучающей выборки составляет всего 114 наблюдений АСМ-изображений с разрешением пикселов 256×256, то, как уже отмечалось, с целью уменьшения числа информативных признаков и, соответственно, исключения «эффек-

та переобучения» частотная область разбивалась на интервалы $\{\Omega_u\}$ по $l = 10$ ($u = 1, 2, \ldots, 7$) точек и для них вычислялись оценки спектральных плотностей $\{\tilde{R}(\omega_k)\}$, используемые в качестве 10 информативных признаков классификации.

Примеры усредненных по всей обучающей выборке спектрограмм для разных типов эритроцитов, вычисленных согласно (1)–(5) по рассматриваемым частотным интервалам, представлены на рисунке.



Усредненные по всей выборке спектрограммы АСМ-изображений поверхностей для класса сфероцитов (——) и для класса дискоцитов (——) (*а*) и фрагменты (при $k = 10, \ldots, 30$) этих спектрограмм (*б*)

При построении дерева решений (6) для каждого интервала доля неклассифицированных наблюдений в терминальных вершинах полагалась равной 0,2. При меньшей доле неклассифицированных наблюдений для анализируемой выборки число терминальных вершин дерева решений и, соответственно, точность классификации обучающей выборки возрастали, но точность классификации экзаменационной выборки при этом уменьшалась.

Минимальная доля ошибок пятикратной кросс-проверки была достигнута при глубине дерева, равной трем. В таблице представлены значения доли правильных решений при классификации по интервалам частот алгоритмами «деревья решений» и «бустинг на деревьях решений».

Точность классификации сфероцитов и дискоцитов по интервалам частот алгоритмами «деревья решений» и «бустинг на деревьях решений», %

Правильные решения	Интервалы частот Ω_u						
	1–10	11–20	21–30	31–40	41–50	51–60	61–70
Деревья решений							
Для выборки сфероцитов	78,43	76,47	70,59	56,86	54,90	72,55	76,47
Для выборки дискоцитов	65,08	79,37	82,54	84,13	80,95	63,44	55,56
При классификации объединенной выборки	71,05	78,07	77,19	73,68	69,30	67,54	65,79
Бустинг на деревьях решений							
Для выборки сфероцитов	76,47	88,24	80,39	80,39	72,55	70,59	72,55
Для выборки дискоцитов	76,47	82,54	79,37	71,42	69,84	68,25	63,44
При классификации объединенной выборки	76,47	85,09	79,84	75,44	71,05	69,30	67,54

Как следует из таблицы, наибольшая точность классификации сфероцитов и дискоцитов алгоритмом «деревья решений» достигается для частотного интервала $\Omega_3 = [21, 30]$. При этом доля ошибочной классификации обучающей выборки составила 0,219, а пятикратной кросс-проверки – 0,267. Доли ошибочной классификации обучающей выборки и кросс-проверки отличаются несущественно. Это свидетельствует о том, что эффект переобучения отсутствует.

При применении алгоритма «бустинг на деревьях решений» объем экзаменационной выборки составил 30 % всей выборки. Объем обучающей выборки наблюдений, использованных для построения каждого дерева, полагался равным 50 % от объема обучающей выборки, которая формировалась с применением процедуры бутстрэпа с возвращением. Значения параметров для бустинга на деревьях решений находились путем перебора по сетке значений таким образом, чтобы ошибки классификации обучающей и экзаменационной выборок были минимальными и сравнимыми между собой.

Заключение. Предложенный метод статистической классификация АСМ-изображений карт физико-механических свойств участков поверхности эритроцитов (дискоцитов и сфероцитов) на основе спектральных признаков обеспечивает достаточно высокую точность классификации – более 82 % (алгоритмом «бустинг на деревьях решений»), что позволяет эффективно использовать его для медико-биологических исследований эритроцитов, полученных от пациентов с наследственным сфероцитозом.

Список использованных источников

1. Dokukin, M. Nanoscale compositional mapping of cells, tissues, and polymers with ringing mode of atomic force microscopy / M. Dokukin, I. Sokolov // Scientific Reports. – 2017. – Vol. 7(1). – P. 11828.

2. Суслов, А. А. Сканирующие зондовые микроскопы (обзор) / А. А. Суслов, С. А. Чижик // Материалы, технологии, инструменты. – 1997. – Т. 2, № 3. – С. 78–89.

3. Temperature- and scale-dependent parameters of lateral force maps of cell surface / M. N. Starodubtseva [et al.] // XIX Annual Linz Winter Workshop. Advances in Single-Molecule Research for Biology & Nanoscience, 3–6 Febr. 2017, Linz, Austria. – Linz, 2017. – P. 6–3.

4. Starodubtseva, M. N. Novel fractal characteristic of atomic force microscopy images / M. N. Starodubtseva, I. E. Starodubtsev, E. G. Starodubtsev // Micron. – 2017. – Vol. 96. – P. 96–102.

5. Kharin, Y. Robustness in Statistical Pattern Recognition / Y. Kharin. – Dordrecht : Kluwer, 1996. – 302 p.

6. Jacobs, T. D. Quantitative characterization of surface topography using spectral analysis / T. D. Jacobs, T. Junge, L. Pastewka // Surface Topography: Metrology and Properties. – 2017. – Vol. 5(1). – P. 013001.

7. Nano- and microscale mechanical properties of erythrocytes in hereditary spherocytosis / M. N. Starodubtseva [et al.] // The Journal of Biomechanics. – 2019. – Vol. 83. – P. 1–8.

8. Андерсон, Т. Статистический анализ временных рядов : пер. с англ. И. Г. Журбенко, В. П. Носко / Т. Андерсон. – М. : Мир, 1976. – 755 с.

9. Harrington, P. Machine Learning in Action / P. Harrington. – N. Y. : Manning, 2012. – 382 p.

10. Hastie, T. The Elements of Statistical Learning / T. Hastie, R. Tibshirani, J. H. Friedman. – 2nd ed. – N. Y. : Springer Publ., 2009. – 764 p.

References

1. Dokukin M., Sokolov I. Nanoscale compositional mapping of cells, tissues, and polymers with ringing mode of atomic force microscopy. *Scientific Reports*, 2017, vol. 7(1), p. 11828.

2. Suslov A., Chizhik S. Skanirujushhie zondovye mikroskopy (obzor) [Scanning probe microscopes (review)]. Materialy, tehnologii, instrumenty [*Materials, Technologies, Tools*], 1997, vol. 2, no. 3, pp. 78–89 (in Russian).

3. Starodubtseva M. N., Yegorenkov N. I., Starodubtsev I. E., Petrenyov D. R., Suslov A. A., Chizhik S. A. Temperature- and scale-dependent parameters of lateral force maps of cell surface. *XIX Annual Linz Winter Workshop. Advances in Single-Molecule Research for Biology & Nanoscience, 3–6 February 2017, Linz, Austria*. Linz, 2017, pp. 6–3.

4. Starodubtseva M. N., Starodubtsev I. E., Starodubtsev E. G. Novel fractal characteristic of atomic force microscopy images. *Micron*, 2017, vol. 96, pp. 96–102.

5. Kharin Y. *Robustness in Statistical Pattern Recognition*. Dordrecht, Kluwer, 1996, 302 p.

6. Jacobs T. D., Junge T., Pastewka L. Quantitative characterization of surface topography using spectral analysis. *Surface Topography: Metrology and Properties*, 2017, vol. 5(1), p. 013001.

7. Starodubtseva M. N., Mitsura E. F., Starodubtsev I. E., Chelnokova I. A., Yegorenkov N. I., Volkova L. I., Kharin Y. S. Nano- and microscale mechanical properties of erythrocytes in hereditary spherocytosis. *The Journal of Biomechanics*, 2019, vol. 83, pp. 1–8.

8. Anderson T. W. *The Statistical Analysis of Time Series.* New York, J. Wiley, 1971, 704 p.

9. Harrington P. *Machine Learning in Action*. New York, Manning, 2012, 382 p.

10. Hastie T., Tibshirani R., Friedman J. H. *The Elements of Statistical Learning*. New York, Springer Publ., 2009, 764 p.

Информация об авторах

Стародубцев Иван Евгеньевич, аспирант, Белорусский государственный университет, Минск, Беларусь.
E-mail: istarodubtsev.science@gmail.com

Харин Юрий Семенович, доктор физико-математических наук, профессор, член-корреспондент НАН Беларуси, Белорусский государственный университет, директор Научно-исследовательского института прикладных проблем математики и информатики Белорусского государственного университета, Минск, Беларусь.

Абрамович Михаил Семенович, кандидат физико-математических наук, доцент, Белорусский государственный университет, заведующий лабораторией статистического анализа и моделирования, Научно-исследовательский институт прикладных проблем математики и информатики Белорусского государственного университета, Минск, Беларусь.

Information about the authors

Ivan E. Starodubtsev, Postgraduate Student, Belarusian State University, Minsk, Belarus.
E-mail: istarodubtsev.science@gmail.com

Yury S. Kharin, Dr. Sci. (Phys.-Math.), Professor, Corresponding Member of the National Academy of Sciences of Belarus, Belarusian State University, Director of Research Institute for Applied Problems of Mathematics and Informatics of the Belarusian State University, Minsk, Belarus.

Mikhail S. Abramovich, Cand. Sci. (Phys.-Math.), Associate Professor, Belarusian State University, Head of Research Laboratory of Statistical Analysis and Modelling, Research Institute for Applied Problems of Mathematics and Informatics of the Belarusian State University, Minsk, Belarus.