УДК 681.3 DOI: 10.33075/2220-5861-2020-1-153-159

СТАТИСТИЧЕСКИЕ МЕТОДЫ КЛАСТЕРИЗАЦИИ ИЗОБРАЖЕНИЙ ГИДРОБИОНТОВ

Ю.Е. Шишкин, А.Н. Греков

Институт природно-технических систем, РФ, г. Севастополь, ул. Ленина, 28
E-mail: iurii.e.shishkin@gmail.com

В работе выполнен анализ эффективности применения статистического подхода к решению задачи кластеризации изображений гидробионтов с использованием модели логистической регрессии для малого числа классов. На примере реальных изображений планктона продемонстрирован процесс построения статистической модели, преобразования изображений отдельных организмов в наборы признаков факторного пространства и построения в нем разделяющих гиперплоскостей. Получена оценка вероятности возникновения ошибок первого и второго рода при осуществлении бинарной кластеризации изображений с использованием разделяющей гиперплоскости.
Ключевые слова: статистическая кластеризация, EM-алгоритм, интеллектуальный анализ данных, машинное обучение, гидробионты, выявление аномалий, логистическая регрессия.

Поступила в редакцию: 14.02.2020. После доработки: 28.03.2020.

Введение. Задача автоматической кластеризации и идентификации видеопотока изображений гидробионтов в реальном масштабе времени исчерпывающе и в полном объеме не разрешена. Решение задачи позволит в значительной степени упростить осуществление численной оценки продуктивности водных экосистем, объема приходящей энергии [1]. Энергия попадает в трофическую сеть, накапливается в виде органических соединений и обеспечивает безостановочное производство биомассы, в частности планктона, который является основой кормового базиса для промысловых рыб [2]. Планктон, несмотря на микроскопические размеры, имеет большую численность, оказывает значительное влияние на процессы, протекающие в морских экосистемах.

В природе ряд планктонных организмов обладает специфичностью к определенным видам физических, химических и радиационных [3] загрязнений что позволяет использовать оценки их популяции в качестве предикторов значимых экологических аномалий, оказывающих значительное влияние не только на хозяйственную деятельность, но и на жизнь человека в целом [4, 5].

Если задача детектирования и отслеживания множественных объектов в видеопотоке является достаточно проработанной и даже имеет ряд успешных практических реализаций, то задача кластеризации и в более широком смысле классификации сопряжена с рядом фундаментальных трудностей [6–8]. В отдельных случаях даже экспертам не всегда удается определить вид планктона по его изображению, не говоря уже о автоматизированных системах классификации.

Решаемая задача в общем виде, по-видимому, на сегодняшний день не имеет конкретного решения ввиду особенностей предметной области: большое разнообразие видов и морфологических признаков планктона, большое внутриклассовое разнообразие, относительное межклассовое сходство. В том случае, когда распознавание происходит вручную, при больших объемах монотонной работы сказывается влияние человеческого фактора.

В статье рассматривается частный случай задачи кластеризации изображений при малом числе кластеров и статистически различимых наборах признаков гидробионтов. Данное допущение справедливо для экосистем с ограниченных видовым разнообразием, например, Черного и Азовского морей.

Постановка задачи. Рассмотрим эффективность применения статистических методов кластеризации изображений гидробионтов, заданных в виде наборов скаляров признакового пространства малой размерности при заданном числе кластеров. А также в общем виде оценим вероятность возникновения ошибки первого и второго рода для бинарного классификатора с целью определения целесообразности применения статистической модели для конкретного рассматриваемого случая.

При использовании классического подхода решения задач распознавания в качестве меры сходства между образами планктона $q \in Q$ используется их расстояние в n-мерном признаковом пространстве Z [9].

Ошибка классификации вызвана несоответствием положения образа $q \in Q$ и соответствующего ему класса X в пространстве $Z \in \Re^n$. Данная ошибка обусловлена стохастической природой объектов исследуемой системы. С другой стороны, в результате применения классификационного алгоритма, может оказаться что образ принадлежит сразу нескольким классам $q \in X_1$ и $q \in X_2$ при этом вероятность этой принадлежности для различных классов неодинакова.

Статистический подход кластеризации изображений гидробионтов как модель логистической регрессии. Статистический подход предусматривает оценку частоты появления образа в каждом из классов. При большом количестве образов (достаточном объеме выборок) в k-м классе эта частота стремится к значению условной вероятности $p_k(q|z \in Z)$ появления образа q в данном классе, где z – множество признаков образа. Под $p_k(q|z)$ будем понимать вероятность того, что образ q, характеризуемый параметрами z, относится к классу X_k. Гистограмму распределения $p_k(q|Z)$ можно рассматривать как дискретную аппроксимацию функции плотности распределения $f_k(q)$, определяющей вероятность того, кто очередной поступивший в систему образ будет принадлежать k-му классу. Таким образом обратная функция $F:Z \rightarrow p_k(q)\ \forall\ k \in X$ будет представлять собой формальное описание модели логистической регрессии, однозначно определяющей вероятность принадлежности образа q, описываемого набором параметров Z, каждому из рассматриваемых классов $X_k\ \forall\ k \in X$.

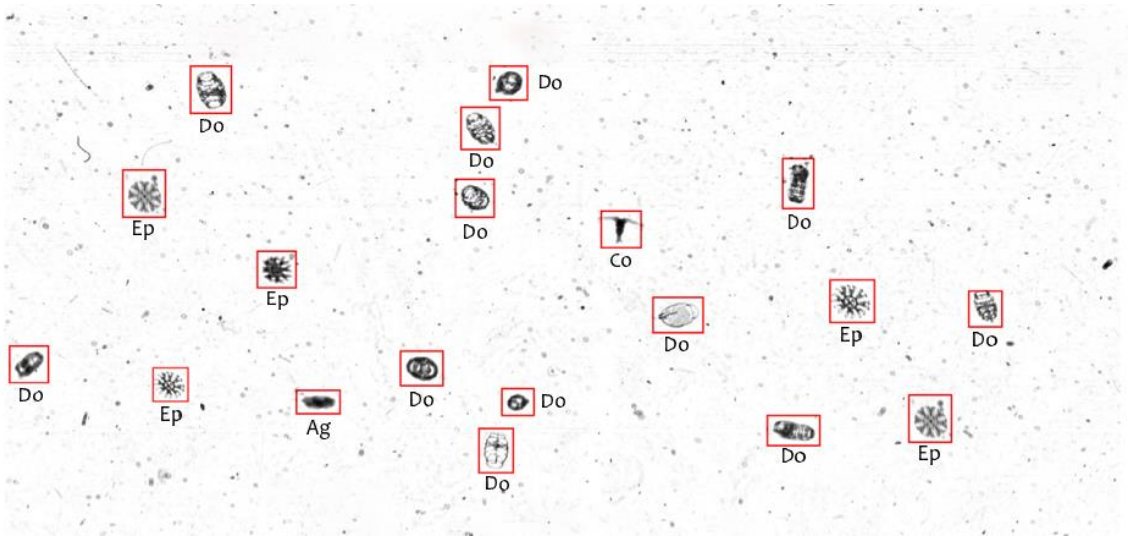

Рис. 1. Изображение исследуемых образцов планктона с разрешением 2048 x 1024 пк. Границы исследуемых объектов обозначены рамками с использованием ISIIS. Объекты промаркированы: Ep – Pelagia noctiluca ephyra (медуза, пелагия), Do – Doliolid (боченочник, долиолид), Co – Calanoid copepod (ракообразный, каляноид), Ag – aggregates (камень, песок) [11]

Fig. 1. Image of studied plankton samples with a resolution of 2048 x 1024 pixels. The boundaries of the studied objects are indicated by frames using ISIIS. Objects are marked: Ep – Pelagia noctiluca ephyra (jellyfish), Do – Doliolid (barrelwort), Co – Calanoid copepod (crustacean), Ag – aggregates (stone, sand)

Пример машинного обучения модели на основе принципа максимума правдоподобия. Существуют решения, позволяющие осуществлять выделение контуров объектов на изображениях [10]. Рассмотрим эффективность применения статистического подхода к кластеризации изображений гидробионтов на примере конкретного снимка с размеченными границами объектов (рис. 1), опубликованного в исследовании [11], полученного с использованием специализированной системы визуализации ихтиопланктона «In situ Ichthyoplankton Imaging System ISIIS».

В математических пакетах обработки данных, решающих задачи распознавания изображений и других прикладных направлений, широко используется набор статистических алгоритмов под общим названием метод максимального правдоподобия ML. Эти алгоритмы представляют собой классификаторы с обучением, в которых исследователями заранее задается конечное множество классов, по которым имеются достаточные объемы обучающих выборок [12, 13]. Эти данные позволяют получить выборки образов по всем классам и оценить вероятность появления образа в каждом классе Ω для всего множества изображений Q.

Сущность применяемого метода для рассматриваемого случая кластеризации изображений гидробионтов состоит в следующем: пусть есть выборки изображений, описываемых соответственно признаками $z_1,\ldots, z_n$, для которых экспертами заданы соответствующие классы Ω. Пусть $L(Z\,|\,\Omega)\colon \Omega \to \Re$ – функция правдоподобия.

Точечная оценка, при использовании ML метода, примет вид

$$ML(z_1,\ldots,z_n) = \arg\max_{z\in Z} L(z_1,\ldots,z_n\,|\,\Omega).$$

Необходимо найти такое правило разделения классов $\Omega_1,\ldots, \Omega_n$, при котором обеспечивается максимизация ML критерия.

Бинарный классификатор изображений планктона на основе вероятностной модели. Построим классификатор для образцов планктона, представленный на рис. 1. Доминирующими объектами на изображении являются Doliolid и Pelagia noctiluca ephyra, обозначим их как Ω_1 и Ω_2 соответственно, поэтому в данном примере рационально использовать бинарный классификатор.

Рассмотрим правило принятия решений с использованием алгоритма статистической классификации на примере двух классов $k = 2$ в двумерном пространстве признаков Z^2, где признаки заданы действительными значениями измерений образа $q \to Z^2$, а именно длинной и шириной объекта.

В рассматриваемом примере осуществляется разделение планктонных организмов классов Ω_1 и Ω_2 по двум признакам x_1 и x_2. В контексте введенных обозначений справедливо следующее $X = \langle \Omega_1, \Omega_2 \rangle$, $Z^2 = \langle x_1, x_2 \rangle$, где Ω_1 соответствует объектам класса Doliolid, Ω_2 – объектам класса Pelagia noctiluca ephyra.

Стохастическая природа метрических характеристик рассматриваемой выборки гидробионтов, изображенной на рис. 1, при их представлении в виде эллипсов, описанных вокруг объектов, может быть задана математической моделью в виде совокупности функций случайных величин, заданных в виде двумерного нормального распределения (1).

Под признаками x_1 и x_2 в данном случае приняты длина и ширина объектов, заданных большой и малой осью эллипсов описанный вокруг каждого гидробионта, однако число признаков в общем виде не ограничивается двумя.

В случае необходимости осуществляется переход в n-мерное признаковое пространство, где $Z^n = \langle x_1,\ldots,x_n \rangle$.

$$f(x_1,x_2,a_1,a_2,\sigma_1,\sigma_2,\sigma_{12}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\sigma_{12}^2}} \exp\left\{ -\frac{1}{2\left(1-\sigma_{12}^2\right)} \times \right.$$
$$\left. \times \left[\frac{(x_1-a_1)^2}{\sigma_1^2} - \frac{2\sigma_{12}(x_1-a_1)(x_2-a_2)}{\sigma_1\sigma_2} + \frac{(x_2-a_2)^2}{\sigma_2^2} \right] \right\}, \quad (1)$$

где σ_{12} – коэффициент парной корреляции между параметрами x_1 и x_2;

a_1, σ_1 – среднее и стандартное отклонение переменной x_1 соответственно;

a_2, σ_2 – среднее и стандартное отклонение переменной x_2 соответственно.

Для иллюстрации предложенного подхода построим математическую модель распределения параметров классов Ω_1 и Ω_2 для ограниченной выборки изображений гидробионтов рис. 1. Будем считать, что случайные величины x_1 и x_2 независимы и взаимная корреляция принимается $\sigma_{12} = 0$, однако в общем случае учет этого коэффициента также необходим.

Для рассмотренного случая математическая модель задается численно совокупностью (1) в виде соотношения

$$F(x_1, x_2) = \begin{cases} f\ (x_1, x_2, 6, 5, 2, 1, 0) \\ f\ (x_1, x_2, 7, 7, 1, 1, 0) \end{cases}. \qquad (2)$$

Статистическая гипотеза в таком случае может быть представлена визуально в виде графиков двух поверхностей (рис. 2), соответствующих классам Ω_1 и Ω_2.

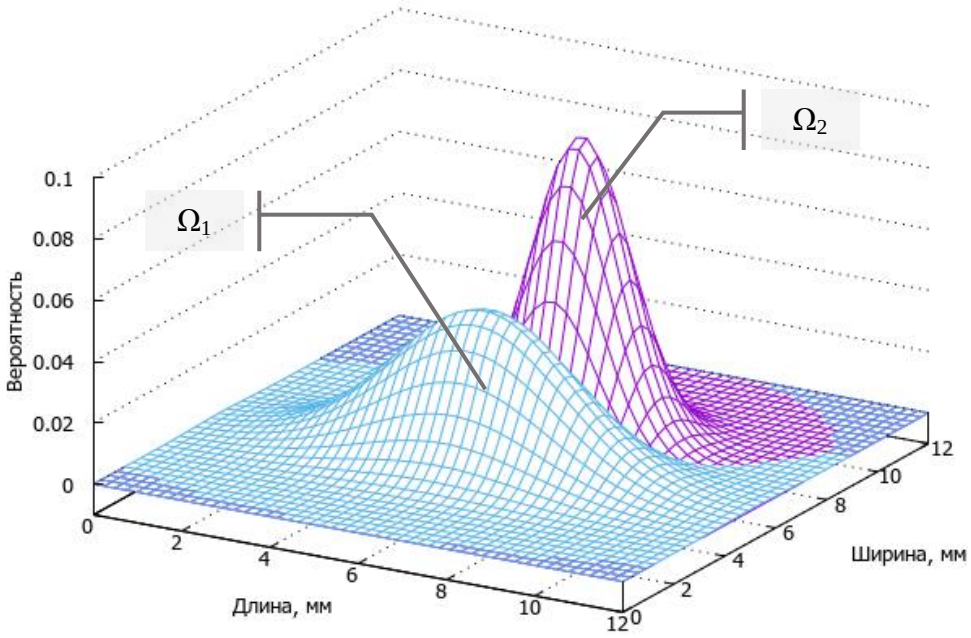

Рис. 2. Статистическая гипотеза для случая двух классов в двумерном пространстве признаков
Fig. 2. Statistical hypothesis for the case of two classes in a two-dimensional space of signs

Задача состоит в определении интервалов Ω_1 и Ω_2, где $\Omega = \langle x_1^{\min}, x_1^{\max}, x_2^{\min}, x_2^{\max} \rangle$, на которых будут приниматься решения в пользу первого и второго класса соответственно. Для простоты в дальнейшем классы и соответствующие им области принятия решения будем обозначать одним и тем же символом Ω.

Допустим, что получен достаточный объем статистической информации о классах Ω, а именно функции плотности статистического распределения $f_1(x_1, x_2)$ и $f_2(x_1, x_2)$, и априорные вероятности $P(x \in \Omega_1|z)$ и $P(x \in \Omega_2|z)\ \forall\ z \in Z$.

Проведем плоскость G $x_2 = k_1 \cdot x_1 + k_2$ разделяющую факторное пространство признаков Z на два подпространства (рис. 3), одно из которых ML: $x_2 < k_1 \cdot x_1 + k_2$ будет соответствовать области решений в пользу класса Ω_1 и $x_2 \geq k_1 \cdot x_1 + k_2$ области решений в пользу класса Ω_2.

Поиск коэффициентов линейной дискриминационной функции ЛДФ k_1 и k_2 обеспечивающих $\alpha = \beta \to \min$ осуществлялся методом наискорейшего спуска.

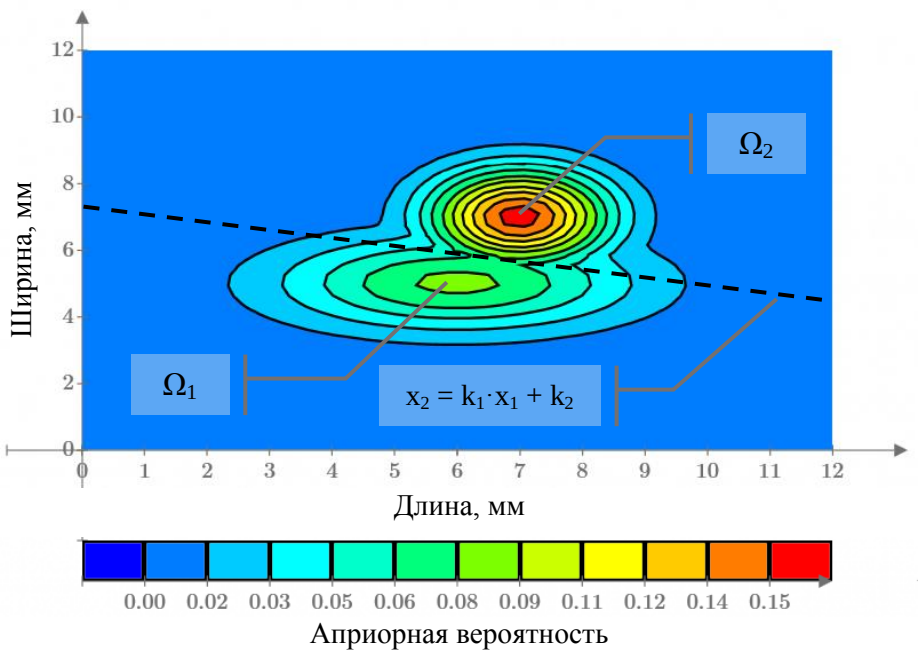

Рис. 3. Контурный график априорной вероятности для случая двух классов в двумерном пространстве признаков и разделяющая их плоскость
Fig. 3. Contour plot of a priori probability for the case of two classes in a two-dimensional space of signs and the straight line dividing them

В общем случае, когда число признаков, описывающих объект интереса равно n, поиск коэффициентов ЛДФ осуществляется в замкнутой области n-мерного пространства любым выбранным исследователем подходящим численным методом. В этом случае, полученные коэффициенты однозначно задают разделяющую гиперплоскость G по аналогии с [14], которая и является границей, определяющей классификационное правило ML.

Оценка вероятности ошибок первого и второго рода при кластеризации. За нулевую гипотезу H_0 примем, что образ планктона с заданными параметрами x_1 и x_2 относится к кластеру Doliolid $ML(x_1, x_2) \to \Omega_1 \mid q \in X_1$.

Альтернативная гипотеза H_1 соответственно $ML(x_1, x_2) \to \Omega_2 \mid q \in X_2$, объект q класса Pelagia noctiluca ephyra.

Ошибка первого рода есть вероятность ошибочно принять гипотезу H_1 когда верна H_0 $\alpha = P(H_1 \mid H_0)$, которая может быть найдена численно как объем, ограниченный графиком поверхности Ω_2, горизонтальной плоскостью, проходящей через начало координат и разделяющей плоскостью G. Аналогично определяется ошибка второго рода $\beta = P(H_0 \mid H_1)$.

Решение задачи поиска объема фигуры с функционально заданными поверхностями тривиально [15] и имеет вид

$$V = \iiint_{\Omega} f(x_1, x_2, z)\,dx_1 dx_2 dz\,.$$

Составим систему ограничений и найдем решение для α и β в общем виде

$$\begin{cases} V = \int_0^T dx_1 \int_0^G dx_2 \int_0^H dF(x_1, x_2) \\ T = -\dfrac{k_2}{k_1} \\ G = k_1 x_1 + k_2 \\ H = \max F(x_1, x_2) \forall x_1, x_2 \in Z^2 \end{cases} . \quad (3)$$

Подставим (1) в (2), а полученный результат в (3) и найдем численное решение для ошибки второго рода

$$\beta = \int_0^T dx_1 \int_0^G dx_2 \int_0^H d \max \begin{cases} \frac{1}{4\pi} e^{-\frac{1}{2}\left(\frac{(x_1-6)^2}{2^2}+\frac{(x_2-5)^2}{1^2}\right)} \\ \frac{1}{2\pi} e^{-\frac{1}{2}\left(\frac{(x_1-7)^2}{1^2}+\frac{(x_2-7)^2}{1^2}\right)} \end{cases}.$$

В рассмотренном примере значения коэффициентов ЛДФ составили $k_1 = -0{,}25$, $k_2 = 7{,}3$, которые задали разделяющую плоскость G.

Графически решение системы (3) проиллюстрировано на рис. 4.

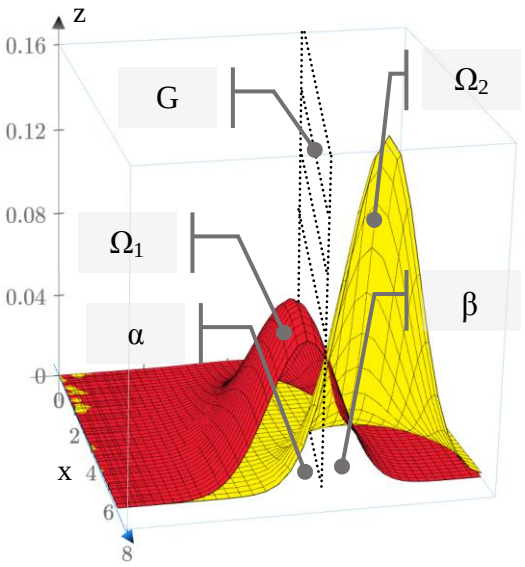

Рис. 4. Сечение графиков поверхностей классов Ω_1 и Ω_2 при $x_1 = 6$ и разделяющая их плоскость G

Fig. 4. The section of Ω_1 and Ω_2 classes surfaces graphs for $x_1 = 6$ and the hyperplane G

Численное решение системы показало, что использование статистического подхода для рассмотренного примера обеспечило одинаковую точность классификации изображений планктона на классы Ω_1 и Ω_2 на уровне $1-\alpha = 1-\beta = 0{,}93$. Вероятности ошибок первого и второго рода составили соответственно $\alpha = \beta = 0{,}07$.

Заключение. При кластеризации изображений гидробионтов, на снимках, полученных с использованием подводной камеры, ошибка первого рода появляется при ошибочном появлении объектов фактически класса Ω_1, среди гидробионтов, классифицированных как посторонние классы. В свою очередь, ошибки второго рода проявляются в ошибочной классификации объектов Ω_1 фактически объектов других классов.

Когда количество классов невелико, обычно преобладают ошибки второго рода. Это связано с тем, что на практике крайне трудоемко учесть все варианты планктона одного вида на изображении, в том числе характеризуемые параметрами, близкими к выделяемым классам. Это одна из причин, по которой целесообразно выполнение предварительной неконтролируемой кластеризации, причем на значительно большее, чем требуется, количество классов, например, для каждой стадии роста объекта. Неконтролируемая классификация позволяет предварительно оценить величину ошибок второго рода, более точно определить границы искомых классов и, при необходимости, разумно задать класс отказов от распознавания.

Следует отметить, модель логистической регрессии, заданная соотношением F, при обоснованном выборе порогового уровня достоверности классификации p_{min}, позволяет также сделать вывод, что условный объект, характеризуемый парой x_1, x_2, не принадлежит ни к одному из имеющихся альтернативных вариантов кластеров Ω. Исследованные модели машинного обучения в совокупности являются основой интеллектуализации при принятии решения о целесообразности применения статистических методов кластеризации для конкретной рассматриваемой задачи.

СПИСОК ЛИТЕРАТУРЫ

1. *Степановских А.С.* Общая экология, М.: Юнити-дана, 2000. 510 с.

2. *Turner J.T.* The importance of small planktonic copepods and their roles in pelagic marine food webs // Zoological Studies, vol. 43, № 2, 2004, С. 255–266.

3. *Thompson P.A.* Plankton: a guide to their ecology and monitoring for water quality // Commonwealth Scientific and

Industrial Research Organization, 2009. С. 7–8. DOI: 10.1071/9780643097131

4. *Шишкин Ю.Е., Скатков А.В.* Информационные технологии обнаружения аномалий в мониторинговых наблюдениях: монография. Симферополь: ИТ «АРИАЛ», 2019. 368 с.

5. *Шишкин Ю.Е., Греков А.Н.* Методы кластеризации изображений для автоматизированного видеорегистратора и анализатора планктона // Комплексные исследования Мирового океана: материалы IV Всерос. науч. конф. молодых ученых. 2019. С. 380–381.

6. *Инзарцев А.В., Павин А.М., Лебедко О.А.* Распознавание и обследование малоразмерных подводных объектов с помощью автономных необитаемых подводных аппаратов // Подводные исследования и робототехника. 2016. № 2 (22). С. 36–43.

7. *Харинов М.В.* Обобщение трех подходов к оптимальной сегментации цифрового изображения // Труды СПИИРАН. 2013. № 2 (25). С. 294–316.

8. *Белим С.В., Кутлунин П.Е.* Выделение контуров на изображениях с помощью алгоритма кластеризации // Компьютерная оптика. 2015. Т. 39. № 1. С. 119–124.

9. *Матвеев Ю.Н.* Основы теории систем и системного анализа, Тверь: ТГТУ, 2007. 100 с.

10. *Shishkin I.E., Grekov A.N.* Analysis of image clusterization methods for oceanographical equipment // 2018 International Russian Automation Conference (RusAutoCon), At Sochi, Russia, September, 2018. DOI: 10.1109/RUSAUTOCON.2018.8501756.

11. *Faillettaz R., Picheral M., Luo J.Y.* Imperfect automatic image classification successfully describes plankton distribution patterns // Methods in Oceanography Vol. 15, 2016. С. 60–77.

12. *Яковлева Т.В.* Условия применимости статистической модели Райса и расчет параметров Райсовского сигнала методом максимума правдоподобия // Компьютерные исследования и моделирование. 2014. Т. 6. № 1. С. 13–25.

13. *Сирота А.А., Соломатин А.И., Воронова Е.В.* Двухэтапный алгоритм обнаружения и оценивания границы объектов на изображениях в условиях аддитивных помех и деформирующих искажений // Компьютерная оптика. 2010. Т. 34. № 1. С. 109–117.

14. *Гданский Н.И., Крашенинников А.М.* Разделение объектов в многомерных пространствах признаков при помощи нормальных классификаторов // Социальная политика и социология. 2012. № 3 (81). С. 202–211.

15. *Туганбаев А.А.* Высшая математика. Функции многих переменных, двойные и тройные интегралы М.: Флинта, 2019 г. 228 с.

STATISTICAL METHODS FOR HYDROBIONT IMAGES CLUSTERING

Iu.E. Shishkin, A.N. Grekov

Institute of Natural and Technical Systems, RF, Sevastopol, Lenin St., 28

The paper analyzes the effectiveness of applying a statistical approach to solving the problem of aquatic organisms images clustering. A logistic regression model for a small number of classes is used. Using real images of plankton as an example, the process of constructing a statistical model transforming images of individual organisms into sets of factor space signs and constructing separating hyperplanes in it is demonstrated. In general, a formula for estimating the first and second kind error probability when performing binary clustering of images using a separating hyperplane has been obtained.
Keywords: statistical clustering, EM algorithm, data mining, machine learning, hydrobionts, anomaly detection, logistic regression.