

© 2022 г. В.В. ЦУРКО, канд. техн. наук (v.tsurko@gmail.com),
А.И. МИХАЛЬСКИЙ, д-р биол. наук, канд. техн. наук (ipuran@yandex.ru)
(Институт проблем управления им. В.А. Трапезникова РАН, Москва)

ОЦЕНКА СТАТИСТИЧЕСКОЙ СВЯЗИ СЛУЧАЙНЫХ ВЕЛИЧИН ЧЕРЕЗ ВЗАИМНУЮ ИНФОРМАЦИЮ

Рассматривается использование непараметрической оценки взаимной информации для выяснения связи между случайными величинами. Показано, что при наличии нелинейной связи между случайными величинами коэффициент корреляции может давать неверный результат. Предложен метод построения оценки взаимной информации по эмпирическим данным в абстрактном гильбертовом пространстве с воспроизводящим ядром. С использованием обобщенной теоремы о представителе предложен метод непараметрической оценки взаимной информации. Работоспособность метода продемонстрирована на примерах анализа искусственных данных. Описано применение метода при прогнозировании устойчивости пентапептидов.

Ключевые слова: коэффициент корреляции, непараметрическая оценка взаимной информации, гильбертово пространство с воспроизводящим ядром, прогнозирование устойчивости пентапептидов.

DOI: 10.31857/S0005231022050063, EDN: ABOPYY

1. Введение

Оценка взаимосвязи случайных величин является одной из центральных задач анализа данных. Наличие статистической зависимости между факторами приводит к плохой обусловленности матрицы плана при построении регрессионных зависимостей [1] и к получению статистически ненадежных оценок влияния этих факторов. При этом часто делается неверный вывод о характере существующей зависимости. Особенно часто такой эффект возникает при анализе медико-биологических данных [2].

Широко используемая мера связи случайных величин — коэффициент корреляции, который эффективен при линейной зависимости. В случае наличия нелинейной зависимости коэффициент корреляции дает неверное значение и даже может свидетельствовать об отсутствии зависимости.

Пример дает пара случайных величин X и Y , связанных через третью случайную величину U , имеющую равномерное распределение на отрезке $[0, 2\pi]$. Положив $X = \sin(U)$ и $Y = \cos(U)$, получим случайные величины, связанные соотношением $X^2 + Y^2 = 1$. Между тем коэффициент корреляции между X

и Y равен нулю:

$$\text{cor}(X, Y) = \frac{1}{2} \int_0^{2\pi} \sin(u) \cos(u) du = \frac{1}{4} \int_0^{2\pi} \sin(2u) du = 0.$$

Другой характеристикой зависимости случайных величин является величина взаимной информации, которую можно интерпретировать как среднее количество информации относительно случайной величины X , содержащейся в случайной величине Y [3, 4]. В [5, 6] рассматриваются вопросы оценивания величины взаимной информации на основании эмпирических данных, обсуждаются трудности и ограничения, возникающие при этом. Использование точных формул для вычисления взаимной информации требует знания совместного распределения случайных величин, оценка которого по эмпирическим данным даже в случае нормального закона распределения требует использования большого числа данных. Та же проблема существует и при попытке оценить по данным непосредственно совместное распределение и применить точные формулы для полученной оценки. В [6] оценка распределения рассматривается как вспомогательная, промежуточная задача и предлагается метод оценки непосредственно взаимной информации.

Отказ от вычисления взаимной информации через оценку совместного распределения рассматривается в [4]. При этом задача ставится как задача минимизации квадратичного функционала, косвенно связанного с взаимной информацией, в многомерном линейном пространстве. Решение этой задачи находится через решение регуляризованной системы линейных алгебраических уравнений.

Предлагаемый в настоящей работе метод оценки взаимной информации также не использует оценку взаимного распределения. Квадратичный функционал, используемый вместо этого, минимизируется в бесконечномерном функциональном гильбертовом пространстве с воспроизводящим ядром. Показано, что решение такой задачи представимо в виде разложения по значениям воспроизводящего ядра в экспериментальных точках. Коэффициенты разложения находятся из регуляризованной системы линейных алгебраических уравнений.

2. Оценка взаимной информации через решение интегрального уравнения

Формально взаимная информация между случайными величинами X и Y , имеющими совместное распределение $P(x, y)$, определяется соотношением

$$I(X, Y) = \int \ln \frac{dP(x, y)}{dP(x) dP(y)} dP(x, y).$$

Если распределение случайных величин имеет плотность $p(x, y)$, то взаимная информация представима в виде

$$I(X, Y) = \int p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} dx dy.$$

Из определения следует, что для независимых случайных величин взаимная информация равна нулю.

Простой метод оценки взаимной информации состоит в замене интегрирования усреднением по выборочным значениям

$$\hat{I}(X, Y) = \frac{1}{km} \sum_{i=1}^k \sum_{j=1}^m p(x_i, y_j) \ln \frac{p(x_i, y_j)}{p(x_i) p(y_j)},$$

где $p(x_i)$, $p(y_j)$ и $p(x_i, y_j)$ – оценки плотностей по экспериментальным данным.

Недостатком прямой замены интегрирования суммированием является требование большого числа экспериментальных данных для получения точных оценок плотностей. Другой подход к оценке взаимной информации без необходимости решения промежуточной задачи – оценки плотностей, предложенный в [6] для случая бинарной классификации. Пусть Y – признак класса, случайная величина, принимающая значения 0 или 1, X – случайный вектор признаков. Выражение для взаимной информации в этом случае записывается в виде

$$I(X, Y) = \sum_{t=\{0,1\}} \int p(x, y=t) \ln \frac{p(x, y=t)}{p(x) p(y=t)} dx.$$

Пусть $x_1^y, \dots, x_{l_y}^y$ – выборка из класса y , l_y – число элементов, принадлежащих классу с меткой y . Запишем эмпирическую оценку взаимной информации в виде

$$I_e(X, Y) = \frac{1}{l_0 + l_1} \sum_{t=\{0,1\}} p(y=t) \left(\sum_{i=1}^{l_0} r_t(x_i^0) \ln r_t(x_i^0) + \sum_{i=1}^{l_1} r_t(x_i^1) \ln r_t(x_i^1) \right),$$

где $r_t(x) = p(x, y=t)/p(x)$ – отношение плотностей. Связь между отношением плотностей и распределением X задается интегральным соотношением

$$F_t(x) = \int \mathbf{I}\{x \geq u\} r_t(u) dF(u),$$

где $F_t(x)$ – функция распределения X в классе с меткой t , $F(x)$ – маргинальная функция распределения X . Данное уравнение можно рассматривать как интегральное уравнение относительно функции $r_t(x)$. Эмпирическим аналогом интегрального уравнения является плохо обусловленная система линейных уравнений

$$\frac{1}{l_t} \sum_{i=1}^{l_t} \mathbf{I}\{x \geq x_i^t\} = \frac{1}{l_0 + l_1} \left(\sum_{i=1}^{l_0} \mathbf{I}\{x \geq x_i^0\} r_t(x_i^0) + \sum_{i=1}^{l_1} \mathbf{I}\{x \geq x_i^1\} r_t(x_i^1) \right),$$

для решения которой в [6] предложен метод регуляризации, использующий специальную V -матрицу, сохраняющую геометрические свойства выборки. При этом, как отмечалось, нет необходимости оценивать плотности распределения вероятностей $p(x)$, $p(x, y)$ и вероятность $p(y)$.

3. Оценка взаимной информации через минимизацию квадратичного функционала

3.1. Метод LSMI

Другая оценка взаимной информации, не требующая предварительной оценки плотностей распределения, предложена в [4]. Отношение плотностей распределения двух случайных величин X и Y

$$w(x, y) = p(x, y) / p(x) / p(y)$$

ищется в виде разложения $\hat{w}(x, y) = \sum_{i=1}^m a_i \varphi_i(x, y)$ по базисным функциям $\varphi_i(x, y)$.

Коэффициенты разложения ищутся путем минимизации функционала

$$J(\hat{w}) = \frac{1}{n^2} \sum_{i,j=1}^n (\hat{w}(x_i, y_j) - 1)^2,$$

где n – число элементов в эмпирической выборке $x_1, y_1, \dots, x_n, y_n$, число членов разложения m является параметром алгоритма. В качестве базисных функций в [4] предлагается использовать гауссианы, за центры которых принимается набор экспериментальных точек. Коэффициенты разложения определяются путем решения регуляризованной системы линейных уравнений. Метод получил название Least-Squares Mutual Information (LSMI). Оценка величины взаимной информации вычисляется как среднее значение оценки $\hat{w}(x, y)$ по эмпирическим данным

$$\hat{I}(X, Y) = \frac{1}{n} \sum_{i=1}^n \ln \hat{w}(x_i, y_i).$$

3.2. Оценка взаимной информации в RKHS

В настоящей работе рассматривается оценка взаимной информации, получающаяся путем минимизации в гильбертовом пространстве с воспроизводящим ядром (RKHS) [7] регуляризованного квадратичного функционала

$$J_\lambda(\hat{w}) = \frac{1}{2} \iint (w(x, y) - \hat{w}(x, y))^2 p(x) p(y) dx dy + \frac{\lambda}{2} \|\hat{w}\|_L^2,$$

где $\|\cdot\|_L$ обозначает норму в гильбертовом пространстве L с воспроизводящим ядром, λ – неотрицательный параметр регуляризации.

Построение оценок случайных процессов в RKHS рассматривалось в [6–8]. Привлекательность гильбертова пространства с воспроизводящим ядром состоит в том, что результат минимизации эмпирического функционала в бесконечномерном гильбертовом пространстве с воспроизводящим ядром находится в конечномерном функциональном пространстве, базисные функции в

котором задаются воспроизводящим ядром гильбертова пространства. В этом заключается суть обобщенной теоремы о представителе [9].

Наиболее нагляден результат использования RKHS в задаче непараметрической оценки регрессионной зависимости по выборке экспериментальных значений $\{x_1, y_1, \dots, x_n, y_n\}$ путем минимизации эмпирического функционала

$$I_{\lambda}(f) = \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_L$$

в гильбертовом пространстве с воспроизводящим ядром $K(x, t)$. В силу обобщенной теоремы о представителе минимум функционала $I_{\lambda}(f)$ достигается на функции вида

$$f(x) = \sum_{i=1}^n a_i K(x, x_i).$$

Коэффициенты разложения a_i являются решением регуляризованной нормальной системы метода наименьших квадратов.

Выбор ядра $K(x, t)$ осуществляется исходя из требуемых свойств решения. Так, если априори известно, что искомая зависимость принадлежит пространству одномерных функций, имеющих ограниченный спектр, то эмпирический функционал надо искать в гильбертовом пространстве с воспроизводящим ядром вида [10]

$$K(x, t) = \frac{\sin(a(x - t))}{\pi(x - t)},$$

где a – граница спектра. Если же искомая зависимость принадлежит пространству одномерных функций, определенных на отрезке $[0, 1]$ и имеющих непрерывную вторую производную, то квадратичный эмпирический функционал надо минимизировать в пространстве кубических полиномиальных сплайнов, которое является гильбертовым пространством с воспроизводящим ядром вида [6]

$$K(x, t) = \int_0^1 (t - \tau)_+ (x - \tau)_+ d\tau = \begin{cases} \frac{1}{2}tx^2 - \frac{1}{6}x^3, & x \leq t \\ \frac{1}{2}xt^2 - \frac{1}{6}t^3, & t \leq x. \end{cases}$$

Чтобы применить обобщенную теорему о представителе для минимизации эмпирической оценки функционала $J_{\lambda}(\hat{w})$, перепишем его в виде

$$J_{\lambda}(\hat{w}) = \frac{1}{2} \iint \hat{w}^2(x, y) p(x) p(y) dx dy - \iint \hat{w}(x, y) p(x, y) dx dy + \\ + \frac{\lambda}{2} \|\hat{w}\|_L^2 + \frac{1}{2} \iint w^2(x, y) p(x) p(y) dx dy.$$

Последнее слагаемое не зависит от $\hat{w}(x, y)$, обозначим его через C . Подставляя вместо пар (x, y) эмпирические значения (x_i, y_i) , $i = 1, \dots, n$, запишем эмпирическую оценку функционала:

$$J_{\lambda}^e(\hat{w}) = \frac{1}{2n^2} \sum_{i=1}^n \sum_{j=1}^n \hat{w}^2(x_i, y_j) - \frac{1}{n} \sum_{i=1}^n \hat{w}(x_i, y_i) + \frac{\lambda}{2} \|\hat{w}\|_L^2 + C.$$

Минимизация $J_{\lambda}^e(\hat{w})$ в гильбертовом пространстве с воспроизводящим ядром $K(z, t)$ приводит к оценке вида

$$\hat{w}(z) = \sum_{i=1}^n \alpha_i K(z, t_i),$$

здесь $z = (x, y)$, $t_i(x_i, y_i)$, коэффициенты α_i ищутся из условия минимизации функционала $J_{\lambda}^e(\hat{w})$ путем решения регуляризованной системы линейных алгебраических уравнений [11]

$$(\mathbf{H} + \lambda \mathbf{K}) \mathbf{a} = \mathbf{h},$$

где элементы матрицы $\mathbf{K}$ вычисляются по формуле

$$K_{ij} = K(x_i, y_i, x_j, y_j),$$

элементы матрицы $\mathbf{H}$ вычисляются по формуле

$$H_{lm} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n K(x_i, y_j, x_l, y_l) K(x_i, y_j, x_m, y_m),$$

$\mathbf{h}$ – вектор с элементами $h_i = \frac{1}{n} \sum_{j=1}^n K(x_i, y_i, x_j, y_j)$.

Для оценки взаимной информации на практике популярным является гауссовское ядро вида $K(z, t) = \exp(-\sigma^{-2} \|z - t\|^2)$. Величины λ и σ являются параметрами метода и выбираются методом кросс-валидации [11].

4. Исследование точности оценки взаимной информации с использованием искусственных данных

Для демонстрации эффективности оценки статистической зависимости через оценку взаимной информации были проведены два статистических эксперимента по сравнению оценки величины взаимной информации с оценкой коэффициента корреляции для пары независимых случайных величин и пары зависимых случайных величин.

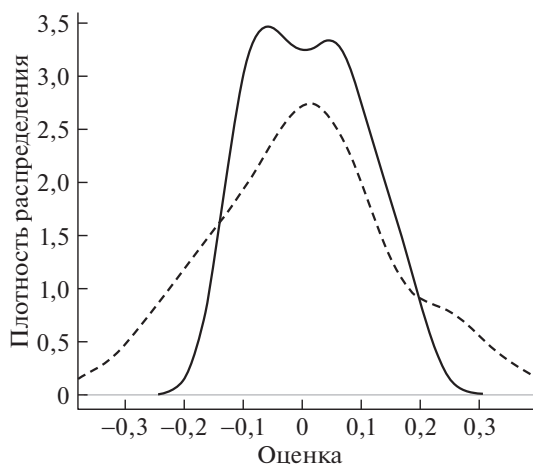


Рис. 1. Плотность распределения оценки взаимной информации двух независимых случайных величин, нормированной на минимальную энтропию (сплошная линия), и оценки по тем же данным коэффициента корреляции (пунктир) при 50 независимых испытаниях.

4.1. Случай независимых случайных величин

Генерировалась независимая выборка из 50 пар независимых нормально распределенных случайных величин X, Y с нулевым средним и единичной дисперсией. По выборке вычислялись оценка коэффициента корреляции и оценка величины взаимной информации. Для нормировки оценка величины взаимной информации делилась на минимальную из оценок для энтропии X и для энтропии Y . Такие выборки независимо генерировались 50 раз. Было получено 50 оценок для коэффициента корреляции и для взаимной информации. Сглаженные распределения оценок представлены на рис. 1.

Из рис. 1 видно, что оценка взаимной информации в среднем ближе к нулю, чем оценка величины коэффициента корреляции.

4.2. Случай зависимых случайных величин

Воспроизводилась описанная выше схема эксперимента с той разницей, что случайные величины X и Y формировались по правилу $X = \sin(U)$, $Y = \cos(U)$, U – случайная величина, равномерно распределенная на $[0, 2\pi]$. Пример реализаций значений X и Y в 50 независимых испытаниях приведен на рис. 2.

Выборки, состоящие из 50 независимо сгенерированных пар X и Y , независимо генерировались 50 раз. Было получено 50 оценок для коэффициента корреляции и для взаимной информации. Сглаженные распределения оценок представлены на рис. 3. Из рис. 3 видно, что оценка взаимной информации существенно отлична от нуля, в то время как оценка величины коэффициента корреляции сосредоточена возле нуля.

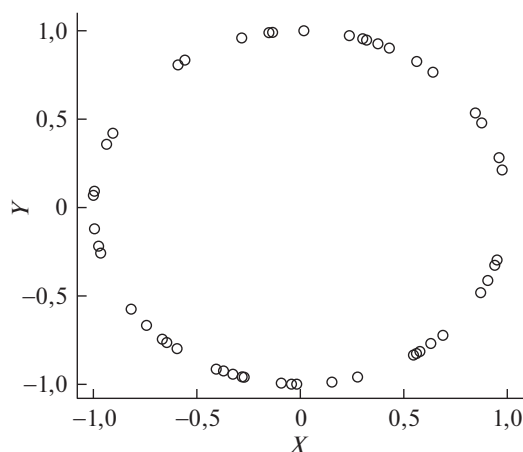


Рис. 2. 50 реализаций значений двух зависимых случайных величин.

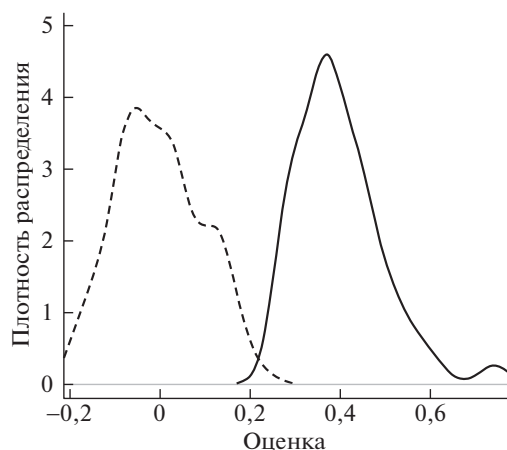


Рис. 3. Плотность распределения оценки взаимной информации двух зависимых случайных величин, нормированной на минимальную энтропию (сплошная линия), и оценки по тем же данным коэффициента корреляции (пунктир) при 50 независимых испытаниях.

В таблице приведены результаты вычисления средних значений и среднеквадратичных отклонений для взаимной информации, нормированной на минимальную энтропию, и трех коэффициентов корреляции Пирсона, Спирмена и Кендалла.

Оценки средних значений и среднеквадратичных отклонений для взаимной информации, коэффициентов корреляции Пирсона, Спирмена, Кендалла

	Взаимная информация	Коэффициент корреляции Пирсона	Коэффициент корреляции Спирмена	Коэффициент корреляции Кендалла
Среднее значение	0,40	-0,0014	-0,0026	-0,00069
Среднеквадратичное отклонение	0,10	0,094	0,076	0,024

5. Применение оценки взаимной информации при оценке стабильности пентапептидов

Описанный в статье метод применялся при решении практической задачи фундаментальной биологии для отбора признаков при прогнозе стабильности конформации коротких белков, состоящих из пяти аминокислот — пентапептидов [11]. Интерес к этой задаче возник после появления публикации [12], в которой исследовалась энтропия белковых последовательностей и было показано, что пентапептиды соответствуют оптимальному размеру структурных блоков, из которых состоят молекулы белков. При увеличении размера структурного блока наблюдался рост информационной энтропии. В [13] методами молекулярной динамики расчетным путем были найдены конформационно-устойчивые пентапептиды и сделано предположение, что такие пептиды играют важную роль в формировании белками нативной пространственной структуры. Такие структурно-устойчивые пептиды могут служить основой формирования элементов вторичной структуры белков и, таким образом, обеспечивать правильную конформацию полипептидной цепи.

Теоретически возможно существование 20^5 разных последовательностей пентапептидов — наборов из пяти аминокислот. Исследование такого количества пентапептидов на устойчивость конформации численными методами молекулярной динамики требует существенного времени и расчетных ресурсов, поэтому прогноз устойчивости конформации пентапептидов методами машинного обучения для последующего изучения методами молекулярной динамики является перспективным.

В [11] для прогноза устойчивости пентапептидов на основании результатов молекулярно-динамического моделирования 49 745 различных пентапептидов применялись два метода классификации: метод ближайших соседей с отбором признаков по величине взаимной информации, оцениваемой в RKHS, как это описано выше, и метод построения случайного леса без предварительного отбора признаков [14]. Результат вычислений показал, что выбор информативных признаков по величине взаимной информации позволил в простом методе ближайших соседей достигнуть качества классификации, близкого к качеству классификации более сложного и ресурсоемкого алгоритма случайного леса, содержащего в себе отбор признаков при построении решающих деревьев.

6. Заключение

Предложенный метод оценки величины взаимной информации случайных величин является существенным дополнением метода оценки коэффициента корреляции. При решении задачи классификации оценка взаимной информации между признаками и меткой класса позволяет находить признаки, связанные с меткой класса не только простыми, близкими к линейным соотношениями, но и сложными нелинейными. Концентрация на таких признаках не только позволяет сократить размерность задачи — число учитываемых признаков, но и учесть влияние признаков, для которых оценка коэффици-

ента корреляции с меткой класса близка к нулю из-за нелинейного характера взаимовлияния.

Основным преимуществом описанного метода является отсутствие необходимости задания априори структуры зависимости между случайными величинами в случае параметрического оценивания либо расположения центров ядерных функций и их числа при непараметрической оценке. Это достигается за счет рассмотрения задачи оценивания взаимной информации в специальном гильбертовом пространстве с воспроизводящим ядром, в котором центры ядерных функций располагаются в экспериментальных точках. Этот факт вытекает из фундаментальной теоремы о представителе [9], следствие которой существенно упрощает построение многих непараметрических оценок.

Для построения непараметрической оценки взаимной информации в статье используется минимизация квадратичного регуляризованного функционала в гильбертовом пространстве с гауссовым воспроизводящим ядром. Параметр ширины ядра выбирался методом кросс-валидации. Приведен симуляционный пример и описана практическая задача из области фундаментальной биологии, демонстрирующие применение предложенного метода оценки взаимной информации.

СПИСОК ЛИТЕРАТУРЫ

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Исследование зависимостей. М.: Финансы и статистика, 1988.
2. Погужев И.Б. Применение математических моделей заболеваний в клинической практике. М.: Наука, 1988.
3. Галлагер Р. Теория информации и надежная связь. М.: Сов. радио, 1974.
4. Suzuki T., Sugiyama M., Kanamori T., and Sese J. Mutual information estimation reveals global associations between stimuli and biological processes // BMC Bioinformatics. 2009. V. 10(Suppl 1):S52.
5. Vergara J.R., Estevez P.A. A review of feature selection methods based on mutual information // Neural Comput & Applic. 2014. V. 24. P. 175–186.
6. Vapnik V., Izmailov R. Statistical inference problems and their rigorous solutions / Gammerman A., Vovk V., Papadopoulos H. (eds) Statistical Learning and Data Sciences. SLDS 2015. Lecture Notes in Computer Science. V. 9047. P. 33–75.
7. Gine E., Nickl R. Mathematical Foundations of Infinite-Dimensional Statistical Model. Cambridge Academ., 2015.
8. Manton J.H., Amblard P.-O. A Primer on Reproducing Kernel Hilbert Spaces // Foundations and Trends in Signal Processing. 2014. V. 8. Is. 1. P. 1–26.
9. Scholkopf B., Herbrich R., and Smola A.J. A generalized representer theorem // Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory, COLT'01/EuroCOLT'01. 2001. P. 416–426.
10. Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Метод потенциальных функций в теории обучения машин. М.: Наука, 1970.

11. *Mikhalskii A.I., Petrov I.V., Tsurko V.V., et al.* Application of mutual information estimation for predicting the structural stability of pentapeptides // RUSS J NUMER ANAL M. 2020. V. 35. No. 5. P. 263–271.
12. *Nekrasov A.* Entropy of Protein Sequences: An Integral Approach // J. Biomolec. Structur. Dynam. 2002. V. 20. P. 87–92.
13. *Nekrasov A., Alekseeva L., Pogosyan R., et al.* A minimum set of stable blocks for rational design of polypeptide chains // Biochimie. 2019. V. 160. P. 88–92.
14. *Breiman L.* Random Forests // Machin. Learning J. 2001. V. 45. No. 1. P. 5–32.

Статья представлена к публикации членом редколлегии А.А. Галяевым.

Поступила в редакцию 12.07.2021

После доработки 03.08.2021

Принята к публикации 26.01.2022