УДК 004.93’11

МЕТОД ИСКУССТВЕННОГО РАЗМНОЖЕНИЯ ДАННЫХ В ЗАДАЧАХ МАШИННОГО ОБУЧЕНИЯ С ИСПОЛЬЗОВАНИЕМ НЕПАРАМЕТРИЧЕСКИХ ЯДЕРНЫХ ОЦЕНОК ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

А. О. Донских, А. А. Сирота

Воронежский государственный университет

Поступила в редакцию 07.08.2017 г.

Аннотация. Рассматривается задача повышения качества обучения алгоритмов классификации образов в условиях недостаточного объема обучающих данных. Предлагается метод искусственного размножения многомерных данных в обучающей выборке на основе процедуры фон Неймана с использованием восстановленной по оригинальным данным многомерной плотности распределения вероятностей вектора признаков. Приводятся результаты сравнения предложенного метода с известными методами решения проблемы ограниченности объема обучающих данных на примере задачи классификации смесей гауссовских случайных величин.
Ключевые слова: искусственное размножение данных, ядерная оценка плотности распределения, процедура фон Неймана, машинное обучение.
Annotation. A problem of increasing classification accuracy of machine learning algorithms on small data sets is considered. A method is proposed to generate additional multivariate training data using nonparametric kernel density estimation and rejection sampling. The method is compared with known methods and techniques for solving the problem of improving small training data classification performance using multivariate Gaussian mixtures as testing data sets. The results of the comparison are provided.
Keywords: data augmentation, kernel density estimation, rejection sampling, machine learning.

ВВЕДЕНИЕ

Одной из проблем, возникающих, при построении алгоритмов машинного обучения и, в частности, алгоритмов классификации (распознавания) образов является формирование исходных обучающих данных. При решении сложных практических задач машинного обучения этап подготовки и организации обучающих данных становится важнейшим, если не сказать самым важным, учитывая общую тенденцию к использованию алгоритмов анализа данных, имеющих универсальную структуру (нейронные сети, композиционные алгоритмы и т. п.).

В то же время поиск объектов, предварительная обработка и формирование их формализованных описаний (признаков) с последующей подготовкой их к использованию для обучения тех или иных алгоритмов требуют значительных затрат времени и человеческих ресурсов. Особенно существенно это проявляется при обработке изображений, звука, текста. По мере усложнения решаемых задач в области анализа данных, создания систем искусственного интеллекта потребности в подобных действиях будут расти, что уже в настоящее время привело к появлению новой области профессиональной деятельности «разметчик данных».

Следует также отметить, что в ряде задач получение представительных обучающих выборок затруднено и по объективным причинам. Обычно это связано с необходимостью постановки и проведения экспериментальных исследований. Такая ситуация возникает, например, при использовании средств спектрального анализа для диагностики па-

тологий биологических объектов (например, элементов зерновых смесей) [1]. При проведении подобных исследований в медицине часто объективно присутствуют ограничения, связанные с необходимостью получения обучающих данных для диагностики заболеваний; при этом подготовка таких данных может занимать месяцы и годы.

К рассматриваемой проблеме имеет также прямое отношение проблема несбалансированности обучающей выборки в задачах обучения алгоритмов классификации. Данная проблема возникает в ситуациях неравномерного представительства различных классов, когда некоторые классы образов представлены существенно меньшими по объему данными, чем другие.

Отсюда возникает потребность в разработке эффективных методов и алгоритмов, обеспечивающих повышение качества машинного обучения и, прежде всего, классификации образов в условиях недостаточного объема обучающих данных.

Одним из перспективных направлений для исследования являются алгоритмы искусственного размножения данных (ИРД) при формировании обучающей выборки. В ряде источников используются схожие, по сути, термины: искусственное расширение, аугментация (от англ. augmentation – приумножение, добавление), морфинг – размножение данных [2-4]. В настоящее время алгоритмы размножения данных применяются преимущественно в задачах обработки изображений. Тем не менее, представляется возможным создание алгоритмов, которые были бы применимы и к числовым данным на основе анализа статистических свойств этих данных по исходной обучающей выборке ограниченного объема.

Важно здесь оговорить, что использование технологий ИРД не может полностью заменить процесс получения представительных реальных обучающих данных, но позволяет, при этом, сократить их объем, а также обеспечить предварительный анализ качества применяемых алгоритмов анализа данных в условиях, когда сбор реальных данных затруднен или требует значительного времени.

Целью данной работы является разработка и исследование метода размножения данных в обучающей выборке на основе восстановления многомерной плотности распределения вероятностей вектора признаков, описывающего объекты в задаче распознавания, с применением ядерных оценок (оценок Парзена-Розенблатта) и генерации новых векторов с использованием метода исключений фон Неймана. Представленный алгоритм имеет и самостоятельное значение: он может использоваться как алгоритм генерации случайных векторов на основе восстановленной многомерной плотности распределения вероятностей по экспериментальным данным.

ИЗВЕСТНЫЕ МЕТОДЫ РЕШЕНИЯ ПРОБЛЕМЫ ОГРАНИЧЕННОСТИ ОБЪЕМА ОБУЧАЮЩИХ ДАННЫХ

Одним из наиболее распространенных способов повышения точности классификации в условиях фиксированного объема обучающей выборки является построение и обучение ансамблей классификаторов. В основе данного подхода лежит идея, что объединение независимых классификаторов в ансамбль позволяет компенсировать их индивидуальные недостатки за счет коллективного голосования, благодаря чему обеспечивается более высокая точность классификации и большая устойчивость к случайным выбросам в обрабатываемых данных. Известные методы построения ансамблей классификаторов можно разделить на зависимые, которые используют классификаторы, полученные на предыдущих этапах, для построения новых, более совершенных классификаторов, и независимые, в которых каждый классификатор ансамбля строится независимо от других [5].

К зависимым методам построения ансамблей классификаторов относят различные реализации подхода, известного как бустинг (boosting) [6], основная идея которого заключается в применении пошаговой процедуры, при реализации которой каждый следующий классификатор стремится компенсировать недостатки композиции всех предыдущих классификаторов, полученной на преды-

дущем шаге. Наиболее известной реализацией бустинга является алгоритм AdaBoost (Adaptive boosting) [7–8].

Среди независимых методов наиболее распространенными являются методы, основанные на бэггинге [9] (bagging, bootstrap aggregating), типичным представителем которых является алгоритм «случайный лес» [10] (random forest). Бэггинг основан на статистическом формирования обучающих данных с использованием бутстрэп-метода [11] (bootstrap), суть которого заключается в формировании множества выборок на основе случайного выбора с повторениями из имеющейся выборки ограниченного объема, что достаточно эффективно имитирует процесс генерации семейства выборок из генеральной совокупности. Полученные методом бутстрэпа выборки используются для обучения независимых классификаторов, которые и образуют ансамбль. В отличие от бустинга, классификаторы в ансамбле не исправляют ошибки друг друга, а компенсируют их при голосовании. Окончательное решение о принадлежности объекта классу может приниматься простым большинством голосов или голосованием с учетом весов отдельных классификаторов ансамбля. Алгоритм «случайный лес» основан на использовании в качестве независимых классификаторов ансамбля деревьев решений, каждое из которых при бэггинге строится по выборке, получаемой из исходной обучающей выборки с помощью бутстрэпа. Использование беггинга, как показывают многочисленные исследования, позволяет снизить требования к объему обучающих данных.

Альтернативный подход к решению проблемы ограниченности объема обучающих данных заключается в реализации прямого искусственного размножения данных на основе исходной обучающей выборки фиксированного объема. Новые данные могут быть получены как путем модификации исходных данных, так и путем генерации новых случайных значений, обладающих какими-либо свойствами, характерными для данных в исходной выборке.

Увеличение объема обучающей выборки за счет трансформаций исходных обучающих данных наиболее часто применяется в задачах распознавания изображений, поэтому данные методы ориентированы преимущественно на обработку изображений. Особенно часто при генерации изображений используются такие преобразования, как поворот на некоторый случайный угол, сжатие и растяжение по вертикали и горизонтали, наклон, зеркальное отражение, обрезка, смещение и многие другие [12–14]. К данной группе методов также можно отнести зашумление исходных данных, а также различные морфинг-преобразования, подобные описанным в работе [4], где новые данные генерируются путем «скрещивания» исходных данных между собой.

Алгоритмы, выполняющие генерацию новых данных, упоминаются в литературе значительно реже бэггинга, бустинга и искажения исходных данных, при этом среди них нельзя выделить некоторые общепризнанные подходы, получившие широкое распространение – как правило, каждая работа содержит описание уникального алгоритма. Тем не менее, данные алгоритмы часто используются в сочетании с методами бустинга и бэггинга, что позволяют получить лучшие результаты.

Так, в работе [15] предложены алгоритмы внесения искусственной реалистической деформации изображений лиц, на основе чего исходная обучающая выборка оригинальных изображений была размножена и использована для обучения алгоритма Виолы-Джонса (алгоритм класса AdaBoost). Показано, что подобным образом объем исходной обучающей выборки может быть уменьшен в 10 раз (1000 оригинальных изображений лиц вместо 10000) при снижении вероятности распознавания не более чем на 2–4 %.

В работе [16] описывается алгоритм, который генерирует значения признаков для искусственного образца как независимые случайные величины, лежащие в диапазоне между минимальным и максимальным значениями соответствующего признака в исходной обучающей выборке. Случайные величины при этом обладают тем же матема-

тическим ожиданием и дисперсией, что и соответствующие признаки исходных данных. В рассматриваемой работе данный алгоритм использовался для искусственного размножения наиболее сложных для распознавания образцов на итерациях модифицированного алгоритма бустинга, предназначенного для решения проблемы несбалансированности обучающей выборки. Полученный в результате алгоритм обеспечил сопоставимую с алгоритмом AdaBoost точность распознавания для относительно сбалансированных наборов данных и значительно более высокую точность для несбалансированных наборов данных.

В работе [17] описывается алгоритм искусственной генерации обучающих данных под названием SMOTE (Synthetic Minority Over-sampling Technique). Идея алгоритма заключается в том, что для каждого опорного образца класса-меньшинства в исходной выборке ищется некоторое количество ближайших соседей. Затем случайным образом выбирается несколько из них, причем количество выбираемых образцов определяется в зависимости от необходимого коэффициента размножения (если объем выборки необходимо увеличить на 200 %, выбирается 2 случайных ближних соседа, если на 300 % – 3 и так далее). Далее для каждого выбранного соседа вычисляется вектор расстояний между его вектором признаков и вектором признаков опорного образца и затем умножается на случайное число в диапазоне от 0 до 1. Полученный вектор суммируется с вектором признаков опорного образца. Таким образом, получается искусственный образ, который располагается в пространстве признаков между рассматриваемым образом и его соседом. Авторы проводят тестирование своего алгоритма на нескольких наборах тестовых данных и отмечают, что в большинстве случаев его применение позволяет получить лучшие результаты по сравнению с традиционной выборкой с повторениями. В своих дальнейших работах [18] авторы объединили бустинг и алгоритм SMOTE, что позволило получить еще более высокие результаты.

ПРЕДЛАГАЕМЫЙ МЕТОД ИСКУССТВЕННОГО РАЗМНОЖЕНИЯ ДАННЫХ

Идея предлагаемого метода ИРД в обучающей выборке базируется на восстановлении многомерной плотности распределения вероятностей вектора признаков, описывающего объекты в задаче распознавания, с применением ядерных оценок и генерации новых векторов с использованием полученной плотности на основе процедуры фон Неймана (метод исключений) [19]. При формировании области исключений должна учитываться специфика использования непараметрической ядерной оценки плотности распределения вероятностей.

Пусть имеется несколько классов образов ω_i, $i=\overline{1,M}$, описываемых случайным вектором признаков распознавания $X=(X_1,...,X_n)^T\in R^n$, представляемого своими значениями в виде вектора $x=(x_1,...,x_n)^T\in R^n$. Для каждого из классов заданы совокупности обучающих данных, как реализации случайного вектора $X=(X_1,...,X_n)^T$ с заданными характеристиками своего класса.

$$X_{le}^{N_i}=\{x_i^{(1)},...,x_i^{(N_i)}\},\ x_i^{(s)}=(x_{i,1}^{(s)},...,x_{i,n}^{(s)})^T\in R^n,$$
$$s=\overline{1,N_i},\ i=\overline{1,M}.$$

На первом этапе реализации предлагаемого метода ИРД для каждого класса проводится восстановление плотности распределения выборки (индекс класса для упрощения записи опускаем). При этом предполагается, что выборка $X_{le}^N=\{x^{(1)},...,x^{(N)}\}$ порождается неизвестным распределением с плотностью $p_X(x/\omega)$. Используется непараметрическая ядерная оценка вида

$$\tilde{p}_X(x/\omega)=\frac{1}{Nh^n}\sum_{s=1}^{N}\varphi\left(\frac{x-x^{(s)}}{h}\right)=$$
$$=\frac{1}{N}\sum_{s=1}^{N}\psi_N\left(x-x^{(s)}\right), \quad (1)$$

где $h^{-n}\varphi\left((x-x^{(s)})/h\right)=K(r),\quad r=\left\|x-x^{(s)}\right\|$ – функция ядра (оконная функция), центрированная относительно каждого вектора обучающей выборки класса и обладающая рядом свойств; h – параметр ширины оконной функции.

Как известно [20], для оценки (1) задаются определенные условия, налагаемые на функцию ядра, а именно:

$$\varphi(u/h)\ge 0,\quad \int \frac{1}{h^n}\varphi(u/h)du=1,$$

$$\int \frac{1}{h^n}\left|\varphi(u/h)\right|du<\infty,\ \ \sup\left|\varphi(u/h)\right|<\infty,$$

которые в совокупности позволяют рассматривать ее как локальную плотность распределения. Обычно функция $\varphi(u/h)$ имеет максимум в точке $u=0$, располагается симметрично относительно этой точки и достаточно быстро убывает при $u\to\infty$. При задании параметра h так, что $h(N)\to 0$, $Nh(N)\to\infty$ при $N\to\infty$, оценка (1) является асимптотически несмещенной и состоятельной.

В качестве универсальной функции ядра, подходящей для многих случаев, целесообразно использовать функцию вида [20–21]

$$\frac{1}{h^n}\varphi\left(\frac{x-x^{(s)}}{h}\right)=\frac{1}{(2\pi)^{n/2}h^n\left|\Xi\right|^{1/2}}\times$$

$$\times\exp\left(-\frac{(x-x^{(s)})^T\Xi^{-1}(x-x^{(s)})}{2h^2}\right)= \qquad (2)$$

$$=N(x,x^{(s)},h^2\Xi),$$

где матрицы Ξ задаются исходя из выборочных оценок матрицы ковариаций класса $\Xi=\tilde{C}$ или в виде одинаковой диагональной матрицы $\Xi=I$.

Ширина оконной функции h существенно влияет на качество восстановления плотности. При $h\to 0$ плотность локализуется вблизи обучающих объектов, а при $h\to\infty$ плотность чрезмерно сглаживается и вырождается в константу. Если распределение образов в признаковом пространстве обладает значительной неравномерностью, то возникает проблема локальных сгущений. Одно и то же значение ширины окна приводит к чрезмерному сглаживанию плотности в одних областях пространства и недостаточному сглаживанию в других. Для решения этой проблемы выбор ширины окна h выполнялся с помощью метода перепроверки на основе принципа максимального правдоподобия [21], который заключается в нахождения максимума логарифма функции правдоподобия при заданной исходной выборке $X_{le}^N=\{x^{(1)},...,x^{(N)}\}$:

$$\ln L(h)\to\max_h, \qquad (3)$$

$$\ln L(h)=\ln\prod_{s=1}^{N}\hat{p}(x^{(s)}/\omega)=$$

$$=\sum_{s=1}^{N}\ln\left(\frac{1}{(N-1)h^n}\sum_{\substack{j=1\\ j\ne s}}^{N}\varphi\left(\frac{x^{(j)}-x^{(s)}}{h}\right)\right).$$

В случае, когда $\varphi(...)$ гауссовская функция вида (2) с матрицей $\Xi=I$, имеем

$$\ln L(h)=\sum_{s=1}^{N}\ln\left(\frac{1}{(N-1)(2\pi)^{n/2}}\frac{1}{h^n}\times\right.$$

$$\left.\times\sum_{\substack{j=1\\ j\ne s}}^{N}\exp\left(-\frac{(x^{(j)}-x^{(s)})^{\mathrm{T}}(x^{(j)}-x^{(s)})}{2h^2}\right)\right). \qquad (4)$$

При этом, как показано в [21], точка максимума $h_0=\arg\max L(h)$ находится в интервале

$$\left[\sqrt{\min_{j\ne s}r_{js}/n},\sqrt{\max_{j\ne s}r_{js}/n}\right],$$

$$r_{js}=(x^{(j)}-x^{(s)})^{\mathrm{T}}(x^{(j)}-x^{(s)}).$$

Таким образом, далее используется оценка $\tilde{p}_X(x/\omega)$ вида (1) с ядром вида (2) с оптимизированным таким образом параметром ядра.

На втором этапе предлагаемого метода ИРД для дальнейшего использования полученной оценки $\tilde{p}_X(x/\omega)$ при генерации реализаций случайных векторов на основе процедуры исключений фон Неймана необходимо определить размеры многомерной области, в пределах которого генерируются равномерно распределенные данные [19]. При этом следует оценить максимум $\max\tilde{p}_X(x/\omega)$ или его верхнюю границу $M_p\ge\tilde{p}_X(x/\omega)$, а также задать границы гиперкуба в пространстве координат $\omega_{x,i}=\{x\ \left|x_{i,\min}\le x_i\le x_{i,\max}\right.\}$, $i=\overline{1,n}$.

Изначально $x_{i,\min}, x_{i,\max}$, $i=\overline{1,n}$ можно установить исходя из значений полученных элементов обучающей выборки, как

$$x_{i,\min}=(1-k_{r,i})\min\{x_i^{(s)}\},$$

$$x_{i,\max}=(1+k_{r,i})\max\{x_i^{(s)}\},\ i=\overline{1,n},$$

где $k_{r,i}$ – некоторый эвристический коэффициент расширения, выбираемый, например, исходя из процентного соотношения отно-

сительного ширины диапазона значений $x_{i,\min} \le x_i \le x_{i,\max}$.

Кроме того, при генерации каждого нового случайного вектора x' при размножении предлагается ввести дополнительное ограничение – пороговое значение для уровня правдоподобия

$$x' \in \Omega_x = \omega_{x,1} \times, ..., \times \omega_{x,n}, \ \tilde{p}_X(x'/\omega) \ge l_0. \quad (5)$$

Следует заметить, что использование более точных и обоснованных границ исходной области $[0, M_p) \times \Omega_x$, а также порога l_0 весьма важно с практической точки зрения, поскольку позволяет сократить время вычислений и общее количество исключений при генерации случайных векторов с заданным законом распределения. Для этого докажем несколько простых утверждений.

Утверждение 1. Пусть $\max\limits_u \varphi(u/h) = \varphi(0)$, тогда для функции $\tilde{p}_X(x/\omega)$ в (1) может быть получена верхняя граница, определяемая следующим неравенством:

$$\max\left[\tilde{p}_X(x/\omega)\right] = \frac{1}{Nh^n}\max\left[\sum_{s=1}^{N}\varphi\left(\frac{x-x^{(s)}}{h}\right)\right] \le$$

$$\le \frac{1}{Nh^n}\sum_{s=1}^{N}\max\left[\varphi\left(\frac{x-x^{(s)}}{h}\right)\right] = \frac{\varphi(0)}{h^n} = M_p. \quad (6)$$

Утверждение 2. Пусть $\max\limits_u \varphi(u/h) = \varphi(0)$ и $\varepsilon > 0$ – вещественное число. Пусть, также $h^{-n}\varphi(u/h) = K(\|u\|) < \delta$ при $\|u/h\| > \varepsilon$ с учетом свойства стремления функции к нулю при $u \to \infty$. Тогда для функции $\tilde{p}_X(x/\omega)$ в (1) может быть определена верхняя граница, достигаемая в точке $\overline{x}$, имеющей в ε-окрестности $\|(x^{(s)} - \overline{x})/h\| \le \varepsilon$ в форме гипершара максимальное число соседей из обучающей выборки $X_{le}^N = \{x^{(1)}, ..., x^{(N)}\}$. Действительно, для любого вектора x выполняется

$$\tilde{p}_X(x/\omega) = \frac{1}{Nh^n}\sum_{s=1}^{N}\varphi\left(\frac{x-x^{(s)}}{h}\right) \le$$

$$\le \frac{N-H(x)}{Nh^n}\delta + \frac{1}{Nh^n}\sum_{s\in S_H}^{H(x)}\varphi\left(\frac{x-x^{(s)}}{h}\right) \le$$

$$\le \frac{N-H(x)}{Nh^n}\delta + \frac{H(x)\varphi(0)}{Nh^n} =$$

$$\frac{N}{Nh^n}\delta + \frac{H(x)(\varphi(0)-\delta)}{Nh^n},$$

где $H(x)$ – количество соседей в ε-окрестности точки x; S_H – множество индексов элементов обучающей выборки, являющихся соседями x.

Отсюда окончательно можно определить точку $\overline{x} = \arg\max H(x)$, для которой выполняется

$$\tilde{p}_X(x/\omega) \le \tilde{p}_X(\overline{x}/\omega) \le$$

$$\le \frac{N}{Nh^n}\delta + \frac{\max\left[H(x)\right](\varphi(0)-\delta)}{Nh^n} \le \quad (7)$$

$$\frac{N}{Nh^n}\delta + \frac{H(\overline{x})\varphi(0)}{Nh^n} = M_p.$$

Сравнение (6) и (7) позволяет сделать вывод о возможности существенного уточнения верхней границы $\tilde{p}_X(\overline{x}/\omega)$. На практике при реализации алгоритма определения верхней границы для $\tilde{p}_X(\overline{x}/\omega)$ можно ограничиться нахождением максимального числа соседей не для любых точек x, хотя бы из ограниченной области, а только для точек $x^{(r)} \in X_{le}^N = \{x^{(1)}, ..., x^{(N)}\}$, являющихся элементами обучающей выборки.

В качестве примера формирования ε-окрестности точки для определения верхней границы рассмотрим ядро вида (2) при $\Xi = I$. Пусть $\|u/h\| \le \varepsilon$, тогда величину δ можно определить следующим образом:

$$\frac{1}{h^n}\varphi\left(\frac{u}{h}\right) = \frac{1}{(2\pi)^{n/2}h^n}\exp\left(-\frac{u^T u}{2h^2}\right) \ge \delta =$$

$$= \frac{1}{(2\pi)^{n/2}h^n}\exp\left(-\frac{\varepsilon^2}{2}\right).$$

Следующее утверждение направлено на анализ влияния порога l_0, ограничивающего снизу оценку функции правдоподобия, с интегральной вероятностью выполнения условия $\tilde{p}_X(x/\omega) \ge l_0$.

Утверждение 3. Пусть область $G = \{x \mid \tilde{p}_X(x/\omega) \ge l_0\}$, $G \subset R^n$ есть область значений x, удовлетворяющих (5). Пусть также $l_0 < N^{-1}\max\limits_u \varphi(u/h) = N^{-1}\varphi(0)$. Определим непустые области $g_s : \{x \mid N^{-1}\psi_N\left(x - x^{(s)}\right) \ge l_0\}$, $s = \overline{1, N}$ и множество $G' = \bigcup\limits_{s=1}^{N} g_s$. Тогда G' является подмножеством G и выполняется следующее неравенство:

$$\int_G \tilde{p}_X(x/\omega)dx \geq L_0' = \int_{g_0} \psi_N(u)\,du,$$

$$g_0 : \{u \mid N^{-1}\psi_N(u) \geq l_0\}. \qquad (8)$$

Действительно, пусть $x \in G' = \bigcup_{s=1}^{N} g_s$, тогда, по крайней мере, для одного из слагаемых ряда (1) в выражении для $\tilde{p}_X(x/\omega)$ выполняется $N^{-1}\psi_N\left(x - x^{(s)}\right) \geq l_0$. Отсюда следует, что и $\tilde{p}_X(x/\omega) \geq l_0$, т. е. $x \in G$ по определению. А это означает, что $G' \subseteq G$.

Из того, что $G' \subseteq G$, а также из свойств функций $\psi_N\left(x - x^{(s)}\right)$, следует цепочка неравенств:

$$L_0 = \int_G \tilde{p}_X(x/\omega)dx \geq \int_{G'} \tilde{p}_X(x/\omega)dx =$$

$$= \frac{1}{N}\sum_{s=1}^{N}\int_{G'} \psi_N\left(x - x^{(s)}\right)dx \geq$$

$$\geq \frac{1}{N}\sum_{s=1}^{N}\int_{g_s} \psi_N\left(x - x^{(s)}\right)dx =$$

$$= \frac{1}{N}\sum_{s=1}^{N}\int_{g_0} \psi_N(u)\,du = \int_{g_0} \psi_N(u)\,du = L_0'.$$

Неравенство (8) позволяет оценить нижнюю границу вероятности L_0 выполнения условия $\tilde{p}_X(x/\omega) \geq l_0$, т. е. степень использования общего объема под поверхностью, восстановленной с помощью непараметрической ядерной оценки плотности распределения при использовании нижнего порога l_0 при реализации метода исключения.

Анализ полученного результата показывает, что значение нижней границы L_0' может быть как ниже значения L_0, так и совпадать с ним. Так, в случае, когда все $x^{(s)} = x^{(*)}$, $s = \overline{1,N}$ равны между собой, плотность локализуется в одной точке и, соответственно, $g_s = g_*$, $s = \overline{1,N}$. Тогда $G' = g_*$, $\tilde{p}_X(x/\omega) = \psi_N\left(x - x^{(*)}\right)$, а, следовательно, пороги, определяющие размеры областей G', G, соответственно, $l_0' = Nl_0$ и l_0, существенно разнятся и, в общем случае, выполняется $\mu(G) \geq \mu(G')$. Это означает, что оценка (8) может быть заниженной, когда обучающие данные локализованы друг относительно друга.

Анализ полученного результата показывает также, что точность полученной оценки нижней границы существенно зависит от вида ядерной функции. Например, если ядро имеет вид гиперкуба, то размеры сечения постоянного уровня функции ядра (размеры областей G', G) не зависят от уровня порога и $\mu(G) = \mu(G')$. Если же размеры сечения постоянного уровня функции ядра существенно зависят от уровня порога (например, для ядерных функций треугольного вида), то $\mu(G) > \mu(G')$ и $L_0 > L_0'$.

Другой случай возникает в ситуации существенного рассредоточения данных, при котором множества g_s, $s = \overline{1,N}$ не пересекаются $g_i \cap g_j = \varnothing$, $i \neq j$. Пусть при этом для каждого x выполняется $\tilde{p}_X(x/\omega) = N^{-1}\psi_N\left(x - x^{(s)}\right)$, где $x^{(s)}$ ближайший к x элемент обучающей выборки. Тогда пороги, определяющие размеры областей G', G равны $l_0' = l_0$, $\mu(G) = \mu(G')$ и в (8) выполняется строгое равенство

$$L_0 = \int_G \tilde{p}_X(x/\omega)dx = \int_{G'} \tilde{p}_X(x/\omega)dx =$$

$$= \frac{1}{N}\sum_{s=1}^{N}\int_{g_s} \psi_N\left(x - x^{(s)}\right)dx = L_0'.$$

Непосредственное вычисление величины L_0, позволяющей использовать неравенство (8), может быть выполнено на основе задания конкретного вида ядра. Так, например, для ядер гауссовского вида задача сводится к расчету объема гиперэллипсоида равной плотности вероятности. Такая задача решалась для случая гауссовской плотности многомерного вектора с независимыми компонентами в [22–24], а, в общем случае, например, в работах [25].

Уравнение гиперэллипсоида может быть получено из условия, определяющего принадлежность точки множеству g_0:

$$\frac{1}{(2\pi)^{n/2}h^n\left|\Xi\right|^{1/2}}\exp\left(-\frac{u^T\Xi^{-1}u}{2h^2}\right) = Nl_0.$$

Отсюда уравнение гиперэллипсоида равной плотности g_0 может быть выражено следующим образом:

$$u^T\Xi^{-1}u = r^2,$$

где $r = h\sqrt{-2\ln\left((2\pi)^{n/2}h^n\left|\Xi\right|^{1/2}Nl_0\right)}$.

Используя полученные в [25] результаты, можно определить нижнюю границу L_0 для

случая гауссовской плотности многомерного вектора на основе следующего выражения:

$$L_0' = P[u \subset g_0] =$$

$$= \left(\int_0^{\infty} t^{n/2-1} e^{-t} dt \right)^{-1} \cdot \int_0^{r^2/2} t^{n/2-1} e^{-t} dt =$$

$$= \frac{\Gamma_{r^2/2}\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = P\left(\frac{n}{2}, \frac{r^2}{2}\right),$$

где $P\left(\frac{n}{2}, \frac{r^2}{2}\right)$ – неполная гамма-функция порядка $r^2/2$ аргумента $n/2$.

Рис. 1. Блок-схема предлагаемого алгоритма искусственного размножения данных

На третьем этапе реализации предлагаемого метода ИРД производится генерация реализаций случайных векторов на основе процедуры исключений фон Неймана с учетом нижнего порога l_0. Процесс генерации новых точек и их отбор продолжается до тех пор, пока не будет получен необходимый объем обучающей выборки.

В итоге алгоритм искусственного размножения данных, реализующий предлагаемый метод, может быть представлен блок-схемой, приведенной на рис. 1.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

Компьютерный эксперимент по исследованию предлагаемого метода искусственного размножения данных производилось с использованием распределений, имеющих вид смесей гауссовских случайных величин (ГСВ). В исходных выборках каждый образ описывается вектором $y = (y_1, ..., y_d)^T \in R^d$, состоящим из $d = 20$ компонент, при этом для обучения и тестирования классификаторов количество признаков сокращалось до $n = 5$ после выполнения линейного преобразования по методу главных компонент $x = (x_1, ..., x_n)^T = Hy$.

Использованная модель гауссовых смесей для каждого класса ω_i, $i = \overline{1, M}$ представляет собой взвешенную сумму m_i компонент и может быть записана выражением:

$$p(y/\omega_i) = \sum_{s=1}^{m_i} p_s^{(i)} b_s^{(i)}(y), \ \sum_{s=1}^{m_i} p_s^{(i)} = 1, \ i = \overline{1, M},$$

где $p_s^{(i)}$ – вероятность появления случайного вектора с параметрами, соответствующими данной компоненте смеси. Обозначим $p^{(i)} = \{p_1^{(i)}, ..., p_{m_i}^{(i)}\}$. При этом каждая компонента $b_s^{(i)}(y)$ является d-мерной гауссовой плотностью распределения вида:

$$b_s^{(i)}(y) = \frac{1}{(2\pi)^{d/2} \left|\Xi_s^{(i)}\right|^{1/2}} \times$$

$$\times \exp\left(-\frac{1}{2} (y - \mu_s^{(i)})^T \Xi_s^{(i)-1} (y - \mu_s^{(i)}) \right).$$

В ходе эксперимента генерировались случайные векторы, соответствующие трем

классов образов, представленных смесями с несколькими компонентами смеси внутри каждого класса (по две компоненты для первого и второго классов и три компоненты для третьего) со следующими параметрами:

$$p^{(1)} = \{0.5, 0.5\},\ p^{(1)} = \{0.5, 0.5\},$$

$$p^{(3)} = \{0.33, 0.33, 0.34\},$$

$$\mu_s^{(i)} = D^{(i)} + rand(0, d\mu),\ D^{(1)} = 0,$$

$$D^{(2)} = 1,\ D^{(3)} = 2,$$

$$\Xi_s^{(i)} = \begin{pmatrix} 1 & r_s^{(i)} & \left(r_s^{(i)}\right)^2 & \dots & \left(r_s^{(i)}\right)^d \\ r_s^{(i)} & 1 & & \dots & \left(r_s^{(i)}\right)^{d-1} \\ \left(r_s^{(i)}\right)^2 & r_s^{(i)} & 1 & \dots & \left(r_s^{(i)}\right)^{d-2} \\ \dots & \dots & \dots & \dots & \dots \\ \left(r_s^{(i)}\right)^d & \left(r_s^{(i)}\right)^{d-1} & \left(r_s^{(i)}\right)^{d-2} & \dots & 1 \end{pmatrix},$$

где $rand(0, d\mu)$ – одномерная равномерно распределенная случайная величина в диапазоне $0...d\mu$; $d\mu$ – параметр, определяющий степень рассредоточенности (пересечения) компонентов смесей каждого класса; $r_s^{(i)} \in [r_{\min}, r_{\max}]$ – коэффициент корреляции, используемый при задании матрицы ковариаций, выбираемый случайным образом в диапазоне $[r_{\min}, r_{\max}]$.

Первоначально проводилась генерация малой обучающей выборки объемом $N = 25$, 50, 75 и 100 образов каждого класса и большой обучающей выборки объемом $N' = 5N = 125$, 250, 375 и 500 образов каждого класса. Эти данные использовались для обучения трех однотипных классификаторов: на основе малой выборки, на основе выборки, полученной путем искусственного размножения малой с увеличением объема данных по каждому классу в 5 раз, а также на основе большой выборки, объем которой получается в этом случае равным искусственно размноженной выборке. Объем тестовых выборок для проверки качества классификации во всех случаях составлял 1000 образов для каждого класса.

В исходных выборках каждый образ описывается $D = 20$ признаками, при этом для обучения и тестирования классификаторов количество признаков сокращалось до $n = 5$ после выполнения линейного преобразования по методу главных компонент.

В качестве обучаемого классификатора для тестирования использовалась нейронная сеть класса MLP (многослойный персептрон). Сеть содержит один скрытый слой с сигмоидальной функцией активации и один выходной слой с линейной функцией активации. Количество входных контактов сети соответствует количеству используемых признаков распознавания $n = 5$, а количество нейронов в выходном слое m_2 равно числу классов (в данном случае 3), при этом значение «1» на выходе нейрона означает, что рассматриваемый образец принадлежит к соответствующему классу, а «0» – не принадлежит. Количество нейронов в скрытом слое m_1 выбиралось из диапазона значений $n \le m_1 \le 2n+1$. Сеть создавалась и тестировалась в среде MATLAB, для обучения использовался алгоритм Левенберга-Марквардта.

Для сравнения эффективности предлагаемого метода с известными методами и алгоритмами в ходе эксперимента были получены результаты для одинаковых наборов обучающих и тестирующих данных для следующих алгоритмов классификации:

- нейронная сеть, обученная по большой выборке;
- нейронная сеть, обученная по малой выборке;
- нейронная сеть, обученная по малой выборке, искусственно размноженной до объема большой выборки с помощью предлагаемого метода;
- нейронная сеть, обученная по малой выборке, искусственно размноженной до объема большой выборки с помощью алгоритма SMOTE [17];
- ансамбль из 10 идентичных по конфигурации нейронных сетей, построенный с помощью традиционного бэггинга с бутстрепом по малой обучающей выборке;
- ансамбль «случайный лес» из 10 деревьев, реализующий бэггинг с бутстрепом по малой обучающей выборке;
- байесовский классификатор, синтезированный для известной плотности распре-

деления данных и используемый для оценки нижней границы вероятности ошибок.

В ходе исследования для каждого набора параметров смесей проводилось по 1000 экспериментов. В ходе каждого эксперимента генерировались новые обучающие и тестовые выборки. Затем полученные результаты усреднялись

На рис. 2 приведены результаты, полученные для смесей ГСВ со слабо коррелированными признаками ($r_{\min} = 0.4,\ r_{\max} = 0.5$). Здесь и далее результаты представлены в виде зависимостей оценки средней вероятности ошибки распознавания классов от объема обучающих данных, задаваемого для малой выборки. При этом результаты работы классификатора, обученного по большой выборке, привязаны к точкам на оси абсцисс исходя из соотношения $N' = 5N$.

Видно, что предлагаемый метод искусственного размножения данных по восстановленной плотности позволяет значительно (до двух раз) сократить вероятность ошибочного распознавания и оказывается эффективнее алгоритма SMOTE. В случае сильно рассредоточенных компонентов смеси $(d\mu = 4)$ ансамбль из 10 нейронных сетей, построенный с помощью процедуры бэггинга, показывает несколько лучшие результаты, что можно объяснить характерной для бэггинга способностью компенсировать неравномерное распределение объектов внутри классов. Ансамбль «случайный лес» из 10 деревьев принятия решений в случае сильно рассредоточенных данных показывает худшие результаты по сравнению с ансамблем нейронных сетей и уступает даже одиночной нейронной сети. Это можно объяснить тем, что нейронная сеть обладает большей обобщающей способностью и лучше адаптируется к подобным данным, чем деревья принятия решений.

На рис. 3 приведены результаты, полученные для смесей ГСВ с сильно коррелированными признаками ($r_{\min} = 0.8,\ r_{\max} = 0.9$). Предлагаемый метод искусственного размножения данных в обоих случаях так же позволяет сократить вероятность ошибочного распознавания, однако в случае сильно рассредоточенных данных $(d\mu = 4)$ алгоритм SMOTE работает чуть лучше, что может быть связано с тем, что основной принцип работы алгоритма SMOTE заключается в заполнении промежутков между данными в обучающей выборке. Ансамбль нейронных сетей, построенный с помощью процедуры бэггинга в случае сильно рассредоточенных компонентов смеси также показывает лучшие результаты.

На рис. 4 приведена зависимость вероятности ошибочного распознавания класса смеси ГСВ со слабо коррелированными при-

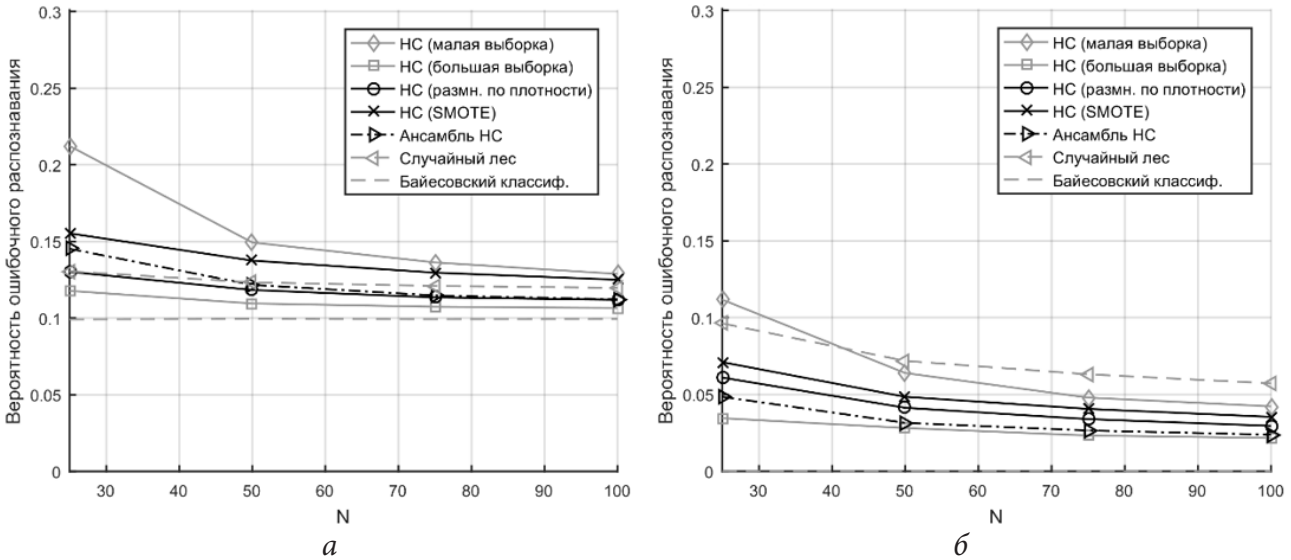
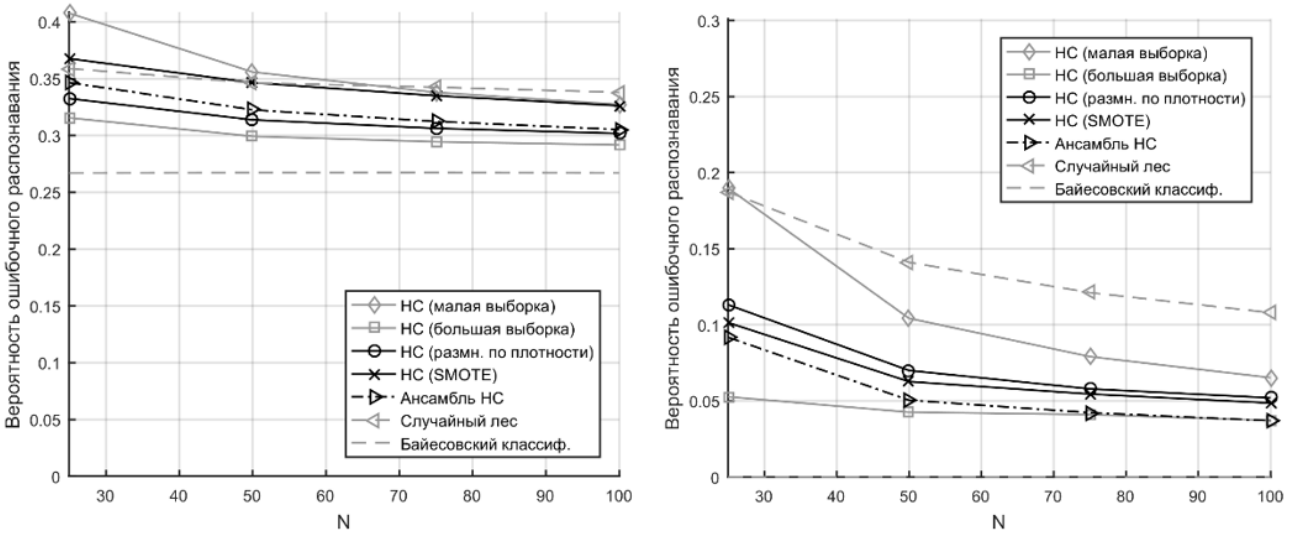


Рис. 2. Зависимость вероятности ошибочного распознавания класса смеси ГСВ при $r_{\min} = 0.4, r_{\max} = 0.5$ *от объема обучающей выборки а) для слабо рассредоточенных данных* $(d\mu = 0)$ *; б) для сильно рассредоточенных данных* $(d\mu = 4)$



а *б*

Рис. 3. Зависимость вероятности ошибочного распознавания класса смеси ГСВ при $r_{\min} = 0.8, r_{\max} = 0.9$ *от объема обучающей выборки а) для слабо рассредоточенных данных* $(d\mu = 0)$*; б) для сильно рассредоточенных данных* $(d\mu = 4)$

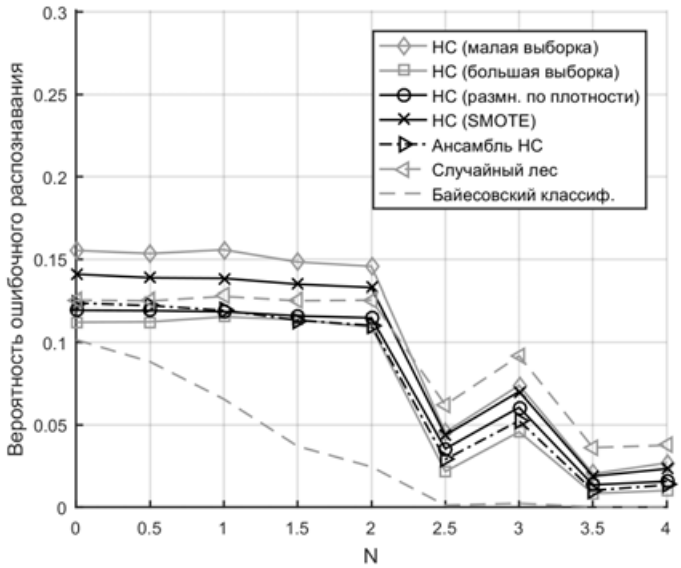


Рис. 4. Зависимость вероятности ошибочного распознавания класса смеси ГСВ при $r_{\min} = 0.4, r_{\max} = 0.5$ *от степени рассредоточенности компонентов смесей* $d\mu$

знаками от степени рассредоточенности компонентов смесей $d\mu$. Видно, что при малых значениях $d\mu$, вероятность ошибочного распознавания не убывает и находится в диапазоне 0.1…0.15. При этом предлагаемый алгоритм размножения позволяет получить преимущество по отношению ко всем другим алгоритмам. При превышении порогового значения $d\mu = 2$ вероятность ошибочного распознавания резко уменьшается и указанное преимущество скрадывается. Это, скорее всего, означает, что применение предлагаемого метода искусственной генерации обучающих данных целесообразно для данных, не имеющих явно выраженных внутренних областей, в которых отсутствует локализация данных («дырок» в многомерном пространстве). Очевидно, что подобная ситуация на практике встречается не часто. В основном, используемые при описании классов образов признаки локализованы, хотя и произвольно, но достаточно компактно.

ЗАКЛЮЧЕНИЕ

В рамках данной работы предложен и исследован метод искусственного размножения многомерных данных в обучающей выборке, основанный на генерации новых реализаций вектора признаков на основе процедуры фон Неймана с использованием восстановленной по оригинальным данным многомерной плотности распределения вероятностей вектора признаков. Метод также может быть использован для моделирования случайных векторов на основе экспериментальных данных произвольной природы. В процессе исследования проведен сравнительный анализ реализованного в рамках предлагаемого метода алгоритма искусственного размножения обучающих данных по отношению к аналогичным алгоритмам такого типа, а также к алгоритмам машинного обучения, основанным на использовании традиционного бэггинга с бутстреп-обработкой данных.

Анализ полученных результатов позволяет сделать вывод о том, что предлагаемый метод позволяет снизить вероятность ошибок классификации при использовании одиночных классификаторов. Предлагаемый метод демонстрирует лучшие среди рассмотренных алгоритмов результаты для данных, не имеющих явно выраженных внутренних областей, в которых отсутствует локализация данных ($d\mu \le 2$). В случае сильно рассредоточенных данных ($d\mu > 2$) алгоритм SMOTE (для сильно коррелированных данных $r_{\min} = 0.8$, $r_{\max} = 0.9$) и ансамбли нейронных сетей, построенных с помощью процедуры бэггинга, оказываются предпочтительнее. Таким образом, одним из перспективных направлений для дальнейшего изучения является исследование возможности комбинирования предложенного метода искусственного размножения данных и традиционных методов построения ансамблей классификаторов.

Результаты работы получены в рамках выполнения государственного задания Минобрнауки России по проекту № 8.3844.2017/4.6 «Разработка средств экспресс-анализа и классификации элементов неоднородного потока зерновых смесей с патологиями на основе интеграции методов спектрального анализа и машинного обучения».

СПИСОК ЛИТЕРАТУРЫ

1. *Донских А. О.* Методы классификации элементов зерновых смесей на основе анализа спектральных характеристик в видимом и инфракрасном диапазонах длин волн / А. О. Донских, Д. А. Минаков, А. А. Сирота, В. А. Шульгин // Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. – 2016. – № 1. – С. 150–160.

2. *Жуковский А. Е.* Синтез обучающей выборки на основе реальных данных в задачах распознавания изображений / А. Е. Жуковский, С. А. Усилин, Н. А. Тарасова, Д. П. Николаев // Информационные технологии и системы (ИТиС'12): сборник трудов конференции. – М., 2012. – С. 377–382.

3. *Ахметов Б. С.* Морфинг-размножение примеров родителей в нескольких поколениях примеров потомков / Б. С. Ахметов, А. И. Иванов, А. Ю. Малыгин, С. В. Качалин, Н. А. Сейлова // Инженерное образование и наука в XXI веке: Проблемы и перспективы: Тр. междун. форума посвящ. 80-летию КазНТУ им. К. И. Сатпаева. – Алматы : КазНТУ, 2014. – Т. 2. – С. 200–203.

4. *Качалин С. В.* Повышение устойчивости обучения больших нейронных сетей дополнением малых обучающих выборок примеров-родителей, синтезированными биометрическими примерами-потомками / Качалин С. В. // Труды научно-технической конференции кластера пензенских предприятий, обеспечивающих безопасность информационных технологий. – Пенза-2014. – Т. 9. – С. 32–35.

5. *Rokach L.* Ensemble-based Classifiers / L. Rokach // Artificial Intelligence Review. – 2010. – V. 33 – P. 1–39.

6. *Schapire R. E.* The strength of weak learnability / R. E. Schapire // Machine Learning. – 1990. – V. 5(2) – P. 197–227.

7. *Freund Y.* Experiments with a new boosting algorithm / Y. Freund, R. E. Schapire // Machine learning: proceedings of the thirteenth international conference. – 1996. – P. 325–332.

8. *Freund Y.* A decision-theoretic generalization of on-line learning and an application to boosting / Y. Freund, R.E. Schapire // Journal of Computer and System Sciences. – 1997. – no.1 – P. 119–139.

9. *Breiman L.* Bagging predictors / L. Breiman // Machine Learning. – 1996. – V. 24(2) – P. 123–140.

10. *Breiman L.* Random forests / L. Breiman // Machine Learning. – 2001. – V. 45(1) – P. 5–32.

11. *Efron B.* Bootstrap Methods: Another Look at the Jackknife / B. Efron // Annals of Statistics. – 1979. – V. 7(1) – P. 1–26.

12. *Yaeger L.* Effective Training of a Neural Network Character Classifier for Word Recognition / L. Yaeger, R. Lyon, B. Webb // NIPS. – 1996.

13. *Ciresan D. C.* Deep Big Simple Neural Nets Excel on Handwritten Digit Recognition / D. C. Ciresan, U. Meier, L. M. Gambardella, J. Schmidhuber // Neural Computation. – 2010. – V. 22(12).

14. *Simard P. Y.* Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis / P. Y. Simard, D. Steinkraus, J. C. Platt // Int'l Conf. Document Analysis and Recognition. – 2003.

15. *Акимов А. В.* Модели и алгоритмы искусственного размножения данных для обучения алгоритмов распознавания лиц методом Виолы – Джонса / А. В. Акимов, А. А. Сирота // Компьютерная оптика. – 2016. – Т. 40, № 6. – С. 911–918.

16. *Guo H.* Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost IM Approach / H. Guo, H. L. Viktor // ACM SIGKDD Explorations Newsletter. – 2004. – V. 6(1). – P. 30–39.

17. *Chawla N. V.* SMOTE: Synthetic Minority Oversampling Technique / N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer // J. Artificial Intelligence Research. – 2002. – V. 16. – P. 321–357.

18. *Chawla N. V.* SMOTEBoost: Improving Prediction of the Minority Class in Boosting / Nitesh V. Chawla, Aleksandar Lazarevic, Lawrence O. Hall and Kevin W. Bowyer // in 7th European Conference on Principles and Practice of Knowledge Discovery in Databases – Cavtat-Dubrovnik, Croatia, September 22–26, 2003. – P. 107–119.

19. *Соболь И. М.* Численные методы Монте-Карло / И. М. Соболь. – М. : ФИЗМАТЛИТ, 1973. – 312 с.

20. *Воронцов К. В.* Математические методы обучения по прецедентам (теория обучения машин) / К. В. Воронцов. – Режим доступа: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf

21. *Кривенко М. П.* Непараметрическое оценивание элементов байесовского классификатора / М. П.Кривенко // Информатика и ее применения. – 2010. – Т. 4, Вып. 2 – С. 13–24.

22. *Вентцель Е. С.* Теория вероятностей. 4-е изд., стереотип / Е. С. Вентцель. – М. : Наука, Физматгиз, 1969 – 576 с.

23. *Лебедев А. М.* Исследование достоверности допускового контроля / А. М. Лебедев // Научный Вестник МГТУ ГА Серия: Эксплуатация воздушного транспорта и ремонт авиационной техники. Безопасность полетов. – 2005. – №86(4). – С. 65–70.

24. *Андронов М. В.* Определение допусков в зависимости от числа контрольных параметров функциональных систем воздушных судов / М. В. Андронов // Научный Вестник МГТУ ГА Серия: Эксплуатация воздушного транспорта. – 2009. – № 147. – С.60–64.

25. *Моисеев А. Н.* Исследование математических моделей систем и сетей массового обслуживания с высокоинтенсивными непуассоновскими входящими потоками: диссертация ... доктора Физико-математических наук: 05.13.18 / Моисеев А. Н.; [Место защиты: Национальный исследовательский Томский государственный университет]. – Томск, 2016. – 333 с.

Донских А. О. – аспирант кафедры Технологий обработки и защиты информации, факультет компьютерных наук, Воронежский государственный университет.
E-mail: a.donskikh@outlook.com

Сирота А. А. – д-р техн. наук, профессор, заведующий кафедрой Технологий обработки и защиты информации, факультет компьютерных наук, Воронежский государственный университет.
E-mail: sir@cs.vsu.ru

Donskikh A. O. – postgraduate student, Department of Processing Technology and Information Security, Computer Sciences Faculty, Voronezh State University.
E-mail: a.donskikh@outlook.com

Sirota A. A. – Doctor of Technical Sciencies, Professor, Head of Department of Processing Technology and Information Security, Computer Sciences Faculty, Voronezh State University.
E-mail: sir@cs.vsu.ru