Смешанные ядерные оценки многомерных распределений и их применение в задачах машинного обучения для классификации биологических объектов на основе спектральных измерений

А.А. Сирота [1], А.О. Донских [1], А.В. Акимов [1], Д.А. Минаков [1]
[1] Воронежский государственный университет, Воронеж, Россия

Аннотация

Рассматривается задача непараметрического восстановления многомерных плотностей распределения вероятностей в системах машинного обучения для классификации и аугментации данных. Предлагается метод получения смешанной ядерной непараметрической оценки плотности распределения как свертки ядерной оценки неизвестной плотности распределения вектора информативных признаков и известной или независимо оцениваемой плотности распределения вектора помеховой составляющей, сопровождающей процесс измерений. Анализируются свойства получаемых таким образом смешанных оценок. Приводятся результаты их сравнения с традиционной ядерной оценкой Парзена, применяемой непосредственно к общей выборке обучающих данных. Теоретически и экспериментально показывается, что использование смешанной оценки эквивалентно реализации процедуры аугментации – искусственного размножения обучающих данных в соответствии с известной или оцененной статистической моделью помеховой составляющей. Рассматриваются возможности применения смешанных оценок для обучения алгоритмов классификации биологических объектов (элементов зерновых смесей) на основе обработки измерений спектров пропускания в видимом и ближнем ИК-диапазонах длин волн.

Ключевые слова: машинное обучение, классификация образов, аугментация данных, ядерная оценка плотности распределения, спектральные измерения.

Цитирование: **Сирота, А.А.** Смешанные ядерные оценки многомерных распределений и их применение в задачах машинного обучения для классификации биологических объектов на основе спектральных измерений / А.А. Сирота, А.О. Донских, А.В. Акимов, Д.А. Минаков // Компьютерная оптика. – 2019. – Т. 43, № 4. – С. 677-691. – DOI: 10.18287/2412-6179-2019-43-4-677-691.

Введение

Во многих задачах анализа данных реализуется подход, основанный на непараметрических ядерных оценках многомерных плотностей распределения вероятностей (ПРВ), которые либо непосредственно используются в алгоритмах классификации объектов, либо обеспечивают решение связанных с ними задач машинного обучения [1–3].

Одной из таких задач, часто применяемых в современных обучающихся системах, является задача аугментации, под которой далее будем понимать искусственное размножение данных (ИРД), т.е. размножение некоторых «опорных» образов на основе стохастических или детерминистских моделей в интересах повышения представительности обучающих выборок. Применение аугментации в условиях малой обучающей выборки во многих случаях позволяет получить более высокую эффективность алгоритмов классификации при их тестировании на новых данных. Данный подход к поиску и подготовке обучающих данных позволяет снизить затраты времени, а также в определенной степени преодолеть проблему несбалансированности обучающих данных различных классов. Следует также отметить, что в некоторых ситуациях сбор необходимого числа обучающих образов оказывается сложен из-за специфического характера предметной области. Как пример реализации ИРД в задаче классификации можно привести предложенный в работе авторов [4] метод размножения данных на основе восстановления многомерной ПРВ вектора признаков из исходной обучающей выборки с применением стандартных ядерных оценок Парзена и генерации новых векторов с использованием метода исключений.

Подход к улучшению качества классификации образов, основанный на ИРД, представлен также в работах [5–12]. Так, в [5–7] для генерации новых изображений в обучающей выборке используются различные преобразования: поворот, сжатие и растяжение, наклон, зеркальное отражение, обрезка, смещение и другие. В [8] новые данные в обучающей выборке генерируются с помощью морфинг-преобразований, путем «скрещивания» исходных данных между собой. В [9] предложены алгоритмы внесения реалистической деформации изображений лиц, с помощью которых была размножена стандартная обучающая выборка изображений для алгоритма Виолы–Джонса. Показано, что подобным образом объем обучающей выборки может быть уменьшен примерно в 10 раз при снижении вероятности распознавания не более чем на 2–4 %.

В [10] описывается алгоритм, который генерирует значения признаков для искусственного образца как независимые случайные величины, лежащие в диапазоне между минимальным и максимальным значениями признака в исходной обучающей выборке. Полу-

ченный алгоритм обеспечил сопоставимую с алгоритмом AdaBoost точность распознавания для относительно сбалансированных наборов данных и значительно более высокую точность для несбалансированных наборов данных. В [11] описывается алгоритм искусственной генерации обучающих данных под названием SMOTE (*Synthetic Minority Oversampling Technique*). Идея алгоритма заключается в том, что для каждого опорного образца класса-меньшинства в исходной выборке случайным образом выбирается некоторое количество ближайших соседей. Далее для каждого выбранного соседа формируется искусственный образ, который располагается случайным образом в пространстве признаков на линии, соединяющей рассматриваемый образ и его соседа. Авторы отмечают, что в большинстве случаев применение алгоритма позволяет получить лучшие результаты по сравнению с традиционной выборкой с повторениями. В своих дальнейших работах [12] авторы объединили бустинг и алгоритм SMOTE, что позволило получить еще более высокие результаты.

В то же время следует отметить, что большинство используемых в указанных работах алгоритмов являются эвристическими, реализующими экспериментально подобранные процедуры зашумления или искажения данных, обеспечивающие получение приемлемого результата в каждом конкретном случае.

Одним из возможных направлений развития методов и алгоритмов аугментации данных является содержательный анализ моделей искажений обрабатываемых сигналов или изображений, обусловленных особенностями проведения измерений признаков, изменением окружающих условий, движением анализируемых объектов и другими факторами, не относящимися к свойствам самих объектов. На практике часто можно говорить о двух составляющих, определяющих статистическую вариативность описания объектов. Первая составляющая является информативной и отражает свойства, характерные для данного класса объектов. Вторая составляющая определяет дополнительные помеховые искажения (ПИ) признаковых описаний объектов, обусловленные перечисленными выше факторами. Наличие ПИ приводит к достаточно большой вариативности регистрируемых признаков классификации, не связанных с вариативностью информативной составляющей. В результате этого для создания качественной обучающей выборки в интересах построения алгоритмов машинного обучения возникает необходимость для каждого образца проводить измерения при всех возможных условиях наблюдения, что во многих случаях связано с большими трудозатратами.

С учетом этих соображений для построения статистических моделей данных и их применения в задачах классификации и аугментации предлагается использовать так называемые смешанные ядерные оценки (СЯО) многомерных плотностей распределения вероятностей, в основе которых лежит независимая оценка ПРВ для вектора информативных признаков и известной или оцененной плотности распределения вектора помеховых искажений, по которым затем вычисляется общая (смешанная) оценка как свертка этих плотностей.

Целью данной работы является анализ свойств смешанных ядерных оценок ПРВ и возможности их применения для классификации и аугментации данных при использовании алгоритмов машинного обучения в интересах обработки спектральных измерений биологических объектов.

1. Метод построения смешанных ядерных оценок и его применение в задачах классификации и аугментации данных

Рассмотрим постановку задачи, согласно которой регистрируемый вектор признаков некоторого класса образов формируется на основе смеси из двух статистически независимых составляющих $X=(X_1,...,X_n)^T=h(U,V)$. Здесь: $U=(U_1,...,U_n)^T$ – случайный вектор, являющийся информативным для решаемой задачи и описываемый характерной для каждого класса образов функцией правдоподобия (ФП), т.е. неизвестным условным распределением $p_U(u/\omega)$, $\omega\in\{\omega_i,\ i=\overline{1,M}\}$; $V=(V_1,...,V_n)^T$ – случайный вектор, не несущий полезную информацию и описываемый единым для всех классов образов распределением $p_V(v/\omega)$. Далее будем использовать обозначения $u=(u_1,...,u_n)^T$, $v=(v_1,...,v_n)^T$ для значений случайных векторов U и V.

В рамках этой постановки можно рассматривать несколько моделей формирования признаков, отличающихся видом оператора $H(U,V)$. Типовой является аддитивная модель. В этом случае выполняется

$$\begin{aligned} &X=U+V,\\ &p_X(x/\omega)=\int p_U(x-y/\omega)\,p_V(y)\,\mathrm{d}y, \end{aligned} \qquad (1)$$

т.е. результирующая плотность распределения формируется как свертка ПРВ составляющих смеси.

Данная модель характеризует действие аддитивных ПИ, линейным образом влияющих на векторное представление каждого образа класса и возникающих в большинстве случаев вследствие потерь при измерении данных (шумов наблюдения, ошибок измерения). Кроме того, аддитивную модель можно использовать для описания естественной изменчивости образов различных классов, вызванной их собственной вариативностью не информативного характера или вариативностью условий регистрации.

Помимо аддитивной модели, может рассматриваться мультипликативная модель ПИ, в которой результирующая плотность распределения формируется как свертка плотностей распределения U и мультипликативной составляющей UV, а также аппликативная модель ПИ, характеризующая действие аномальных ошибок, приводящих к потере части полезной информации, когда отдельные компоненты вектора u, отражающего свойства образов класса, замещаются компонентами вектора v, никак с этим классом не связанного.

Далее будем рассматривать аддитивную модель. Для получения оценок в рамках этой модели возможны различные ситуации, характеризующиеся разным уровнем априорной неопределенности.

1.1. Смешанные ядерные оценки и их свойства

Пусть при построении распределения смеси $p_X(x/\omega)$ ПРВ $p_U(u/\omega)$ неизвестна, а ПРВ $p_V(v)$ известна или заранее независимо оценена любым из возможных способов и характеризует ПИ, воздействующие на исходные образы класса в соответствии с (1). При этом для построения плотности $p_U(u/\omega)$ задана обучающая выборка в виде совокупности реализаций $U_{le}{}^N=\{u^{(1)},...,u^{(N)}\}$. Тогда получение искомой оценки ФП предлагается проводить в два этапа. На первом этапе на основе обучающей выборки может быть получена оценка

$$\tilde{p}_U(u/\omega)=\frac{1}{Nh^n}\sum_{s=1}^{N}\varphi\left(\frac{u-u^{(s)}}{h}\right)=\\ =\frac{1}{N}\sum_{s=1}^{N}\psi_N\left(u-u^{(s)}\right), \tag{2}$$

где $\psi_N(u-u^{(s)})=h^{-n}\phi((u-u^{(s)})/h)$ – функция ядра (оконной функции); h – параметр ядра (ширина оконной функции).

На втором этапе осуществляется свертка полученной оценки распределения (2) $\tilde{p}_U(u/\omega)$ с учётом $U=X-V$ и распределения аддитивного шума в соответствии с соотношением:

$$\tilde{\tilde{p}}_X(x/\omega)=\int\tilde{p}_U(x-v/\omega)p_V(v)\mathrm{d}v=\\ =\frac{1}{Nh^n}\sum_{s=1}^{N}\int\varphi\left(\frac{(x-v)-u^{(s)}}{h}\right)p_V(v)\mathrm{d}v=\\ =\frac{1}{Nh^n}\sum_{s=1}^{N}\int\varphi\left(\frac{z-u^{(s)}}{h}\right)p_V(x-z)\mathrm{d}z=\\ =\frac{1}{N}\sum_{s=1}^{N}\vartheta_N\left(x-u^{(s)}\right),\\ \vartheta_N\left(x-u^{(s)}\right)=\int\frac{1}{h^n}\varphi\left(\frac{x-v-u^{(s)}}{h}\right)p_V(v)\,\mathrm{d}v=\\ =\int\psi_N\left(x-u^{(s)}-v\right)p_V(v)\,\mathrm{d}v, \tag{3}$$

т.е. определяется конечной суммой, элементы которой являются свертками ядерных функций и плотности $p_V(v)$, центрированных относительно точек совокупности $U_{le}{}^N=\{u^{(1)},...,u^{(N)}\}$. Например, когда используются ядра в виде гауссианы $h^{-n}\varphi((u-u^{(s)})/h)=N(u,u^{(s)},h^2\,\Xi)$ и присутствует аддитивный шум, имеющий центрированное гауссовское распределение с матрицей ковариации C_v $p(v)=N(v,0,C_v)$, после свертки получается оценка в виде

$$\tilde{\tilde{p}}_X(x/\omega)=\frac{1}{N}\sum_{k=1}^{N}\mathrm{N}(x,u^{(s)},h^2\Xi\ +C_v). \tag{4}$$

Полученные таким образом оценки будем назвать смешанными, акцентируя внимание на том, что одна из составляющих смеси получена на основе непараметрической оценки, а другая определена изначально (задана в аналитическом виде).

Анализ несмещённости и состоятельности оценки

Как известно [13–15], непосредственные оценки Парзена на основе соотношений вида (2) при определенных условиях являются несмещенными и состоятельными. Пусть для оценки (2) выполняются следующие свойства функции ядра

$$\int\frac{1}{h^n}\varphi(u/h)\,\mathrm{d}u=\int\varphi(z)\,\mathrm{d}z=1,\\ \int\frac{1}{h^n}\left|\varphi(u/h)\right|\mathrm{d}u=\int\left|\varphi(z)\right|\mathrm{d}z<\infty,\\ \sup\left|\varphi(u/h)\right|=\sup\left|\varphi(z)\right|<\infty. \tag{5}$$

Указанные свойства свидетельствуют о том, что функция ядра удовлетворяет свойствам плотности распределения. Пусть также при задании параметра ядра как функции наблюдений $h=h(N)$ и $\psi_N(u)=h^{-n}(N)\varphi(u/h(N))$ [15] для каждой точки непрерывности плотности $\tilde{p}_U(u/\omega)$ выполняется

$$\int\psi_N(u-y)p_U(y/\omega)\mathrm{d}y\underset{N\to\infty}{\to}p_U(u/\omega),\\ \sup\frac{1}{N}\psi_N(u)\underset{N\to\infty}{\to}0.$$

Тогда оценка (2) является асимптотически несмещённой и состоятельной [15]. Отметим, что первое из этих условий означает, что последовательность функций $\{\psi_N(u-y)\}$ при $N\to\infty$ сходится к дельта – функции с центром в точке y. Для широкого класса функций, которые могут быть использованы в качестве ядра и обладают свойствами (5), такая сходимость имеет место, если при $N\to\infty$ $h(N)\to 0$. Второе свойство (состоятельности) для таких функций выполняется, если при $N\to\infty$ $Nh^n(N)\to\infty$ [15].

Асимптотическая несмещенность оценки (2) по совокупности $U_{le}{}^N=\{u^{(1)},...,u^{(N)}\}$ может быть показана на основе следующих преобразований [15]:

$$M\left[\tilde{p}_U(u/\omega)\right]=\frac{1}{Nh^n}\sum_{s=1}^{N}M\left[\varphi\left(\frac{u-U^{(s)}}{h}\right)\right]=\\ =\frac{1}{h^n}M\left[\varphi\left(\frac{u-U}{h}\right)\right]=\\ =\int\frac{1}{h^n}\varphi\left(\frac{u-y}{h}\right)p_U(y/\omega)\,\mathrm{d}y=\\ =\int\psi_N(u-y)p_U(y/\omega)\,\mathrm{d}y\underset{N\to\infty}{\to}p_U(u/\omega). \tag{6}$$

Утверждение 1. Если $\tilde{p}_U(u/\omega)$ при $N\to\infty$ $h(N)\to 0$ является асимптотически несмещённой оценкой ПРВ $p(u/\omega)$ для всех u, то смешанная оценка $\tilde{\tilde{p}}(x/\omega)$ является асимптотически несмещённой оценкой $p(x/\omega)$ [16].

Для доказательства рассмотрим выборку $X_{le}{}^N=\{x^{(1)},...,x^{(N)}\}$, где каждый $x^{(s)}=u^{(s)}+v^{(s)}$. По аналогии с (6) для стандартной оценки выполняется

$$M\left[\tilde{p}_X(x/\omega)\right]=\int\frac{1}{h^n}\varphi\left(\frac{x-y}{h}\right)p_X(y/\omega)\,\mathrm{d}y=$$
$$=\int\psi_N(x-y)\,p_X(y/\omega)\,\mathrm{d}y\underset{N\to\infty}{\to}p_X(x/\omega).$$

С другой стороны

$$M\left[\tilde{p}_X(x/\omega)\right]=\frac{1}{Nh^n}\sum_{s=1}^{N}M\left[\varphi\left(\frac{x-U^{(s)}-V^{(s)}}{h}\right)\right]=$$
$$=\frac{1}{h^n}M\left[\varphi\left(\frac{x-U-V}{h}\right)\right]=$$
$$=\int\frac{1}{h^n}\varphi\left(\frac{x-u-v}{h}\right)p_U(u)\,p_V(v)\,\mathrm{d}u\,\mathrm{d}v=$$
$$=\int\vartheta_N(x-u)\,p_U(u)\,\mathrm{d}u\,.$$

Последняя запись использует функцию, суммируемую в представлении смешанной оценки (3), и, следовательно, стандартная оценка имеет то же математическое ожидание, что и смешанная, т.е. для неё также выполняется свойство асимптотической несмещённости.

Как известно, если оценка сходится к истинному значению параметра «в среднем квадратичном» или если оценка асимптотически несмещённая и её дисперсия стремится к нулю, то такая оценка будет состоятельной. Таким образом, состоятельность стандартной оценки (2) вытекает из её сходимости в среднеквадратичном [13, 15], т. е. при $N\to\infty$

$$D\left[\tilde{p}_U(u/\omega)\right]=M[\tilde{p}_U^2(u/\omega)]-p_U^2(u/\omega)\to 0.$$

Утверждение 2. Если $\tilde{p}_U(u/\omega)$ при $N\to\infty$ $h(N)\to 0$ сходится в среднеквадратичном к ПРВ $p(u/\omega)$ для всех u, то смешанная оценка $\tilde{\tilde{p}}(x/\omega)$ аналогичным образом сходится к $p(x/\omega)$ [16].

Рассмотрим сначала дисперсию стандартной оценки

$$D\left[\tilde{p}_X(x/\omega)\right]=M[\tilde{p}_X^2(x/\omega)]-p_X^2(x/\omega)$$

и представим в ней математическое ожидание квадрата оценки в виде

$$M[\tilde{p}_X^2(x/\omega)]=$$
$$=\frac{1}{N^2h^{2n}}\sum_{r=1}^{N}\sum_{s=1}^{N}M\left[\varphi\left(\frac{x-X^{(s)}}{h}\right)\varphi\left(\frac{x-X^{(r)}}{h}\right)\right]=$$
$$=\frac{1}{N^2}\sum_{s=1}^{N}M\left[\psi_N^2\left(x-X^{(s)}\right)\right]+$$
$$+\frac{1}{N^2}\sum_{r=1}^{N}\sum_{\substack{s=1,\\ s\neq r}}^{N}M\left[\psi_N\left(x-X^{(s)}\right)\psi_N\left(x-X^{(r)}\right)\right]=$$
$$=\frac{1}{N}M\left[\psi_N^2\left(x-X\right)\right]+$$
$$+\frac{N-1}{N}M\left[\psi_N\left(x-X\right)\right]^2\underset{N\to\infty}{\to}p_X^2(x/\omega).$$

Здесь для первого слагаемого при

$$\sup(1/N)\psi_N(x)\underset{N\to\infty}{\to}0\ \ (h(N)\to 0,\ Nh^n(\mathrm{N})\to\infty)$$

выполняется

$$\frac{1}{N}\int\psi_N^2\left(x-u-v\right)p_U(u)\,p_V(v)\,\mathrm{d}u\,\mathrm{d}v\le$$
$$\le\frac{1}{N}\sup\psi_N(x)\int\psi_N\left(x-u-v\right)p_U(u)\,p_V(v)\,\mathrm{d}u\,\mathrm{d}v\to 0.$$

Для второго слагаемого при $\sup(1/N)\psi_N(x)\underset{N\to\infty}{\to}0$ $(h(N)\to 0,\ Nh^n(N)\to\infty)$ в силу сходимости стандартных оценок выполняется

$$\frac{N-1}{N}\left[\int\psi_N\left(x-u-v\right)p_U(u)\,p_V(v)\,\mathrm{d}u\,\mathrm{d}v\right]^2=$$
$$=\frac{N-1}{N}\left[\int\vartheta_N\left(x-u\right)p_U(u)\,\mathrm{d}u\right]^2\underset{N\to\infty}{\to}p_X^2(x/\omega).$$

Теперь рассмотрим дисперсию смешанной оценки

$$D\left[\tilde{\tilde{p}}_X(x/\omega)\right]=M[\tilde{\tilde{p}}_X^2(x/\omega)]-p_X^2(x/\omega).$$

Заметим сначала, что

$$\frac{1}{N}\vartheta_N(x)=\frac{1}{N}\int\psi_N\left(x-v\right)p_V(v)\,\mathrm{d}v\le\frac{1}{N}\sup\psi_N(x).$$

Тогда для смешанной оценки

$$M[\tilde{\tilde{p}}_X^2(x/\omega)]=$$
$$=\frac{1}{N^2}\sum_{r=1}^{N}\sum_{s=1}^{N}M\left[\vartheta_N\left(x-U^{(s)}\right)\vartheta_N\left(x-U^{(r)}\right)\right]=$$
$$=\frac{1}{N^2}\sum_{s=1}^{N}M\left[\vartheta_N^2\left(x-U^{(s)}\right)\right]+$$
$$+\frac{1}{N^2}\sum_{r=1}^{N}\sum_{\substack{s=1,\\ s\neq r}}^{N}M\left[\vartheta_N\left(x-U^{(s)}\right)\vartheta_N\left(x-U^{(r)}\right)\right]=$$
$$=\frac{1}{N}M\left[\vartheta_N^2\left(x-U\right)\right]+\frac{N-1}{N}M\left[\vartheta_N\left(x-U\right)\right]^2.$$

При этом первое слагаемое при $\sup(1/N)\psi_N(x)\underset{N\to\infty}{\to}0$ стремится к нулю, так как

$$(1/N)\int\vartheta_N^2\left(x-u\right)p_U(u)\,\mathrm{d}u\le$$
$$\le(1/N)\sup\vartheta_N(x)\int\vartheta_N\left(x-u\right)p_U(u/\omega)\,\mathrm{d}u\underset{N\to\infty}{\to}0,$$

а для второго слагаемого выполняется

$$\frac{N-1}{N}M\left[\vartheta_N\left(x-U\right)\right]^2=\frac{N-1}{N}\left[\int\vartheta_N\left(x-u\right)p_U(u/\omega)\,\mathrm{d}u\right]^2,$$

т.е. математические ожидания квадрата стандартной и смешанной оценок стремятся к одному пределу, определяемому величиной $p_X^2(x/\omega)$. Это позволяет говорить о сходимости при $h(N)\to 0$ смешанной оценки в среднеквадратичном $D[\tilde{\tilde{p}}_X(x/\omega)]\to\infty$ и о её состоятельности в целом. Проведённый анализ позволяет говорить о возможности применения СЯО, наряду со стандартными непараметрическими оценками (оценками Парзена).

1.2. Использование моделей смешанных ядерных оценок для искусственного размножения данных

Утверждение 3. Пусть $U_{le}{}^N=\{u^{(1)},\ldots,u^{(N)}\}$, где каждый $u^{(s)}$ является реализацией U. Пусть при за-

данной ПРВ $p(v)$, $M[v]=0$ для каждого вектора $u^{(s)} \in U_{le}{}^{N}$ проводится искусственное размножение данных путём генерации подвыборки, состоящей из P образов, на основе распределения $p_{W_s}(w/u^{(s)})$, где $w=u^{(s)}+v+g$, причём g – случайная величина с ПРВ, определяемой одним из возможных ядер, используемых при получении смешанной оценки, $p_G(g)=h'^{-n}\varphi(g/h')=\psi_N(g)$, $M[g]=0$. Здесь $h'=(N)$ – значение, выбираемое исходя из заданного N.

Тогда при $P\to\infty$ реализация подобной процедуры размножения обучающих данных позволяет восстановить плотность распределения $\tilde{\tilde{p}}_X(x/\omega)$ смешанной оценки, что даёт возможность говорить об эквивалентности применения рассматриваемой процедуры ИРД и СЯО для генерации данных при обучении алгоритмов классификации.

Для генерации данных на основе ПРВ известной $p_V(v)$, $M[v]=0$ может быть использован любой известный метод моделирования случайных векторов, например метод исключений. Отметим также, что плотность распределения $p_{W_s}(w/u^{(s)})$ $w=u^{(s)}+q$ для любой q, если известна $p_Q(q)$, определяется выражением $p_{W_s}(w/u^{(s)})=p_Q(w-u^{(s)})$.

Для доказательства рассмотрим выборку данных следующего вида

$$UV_{le}^{N}=\{U_{le+}^{(1)},...,U_{le+}^{(N)}\},$$
$$U_{le+}^{(s)}=\{w^{(s,1)},...,w^{(s,P)}\},$$
$$w^{(s,k)}=u^{(s)}+v^{(s,k)}+g^{(s,k)},k=\overline{1,P},$$

т.е. выборку данных, полученных при генерации реализаций вектора аддитивного шума $q=v+g$ на основе ПРВ $p(v)$, $M[v]=0$, $p_G(g)$, $M[g]=0$, добавляемых к точкам $u^{(s)}$ и сгруппированных в соответствующие подвыборки.

Рассмотрим теперь ядерную оценку ПРВ $\tilde{p}_W(w/\omega)$ для величины $w=u+v+g=u+q$ по выборке $UV_{le}{}^{(N)}$, в которой параметр $h(PN)=h(N_{sum})$ выбирается в соответствии с исходными свойствами при $N_{sum}\to\infty$ $h(N_{sum})\to 0$ и $N_{sum}h^n(N_{sum})\to\infty$. Очевидно, что в случае, если $P\to\infty$, то $N_{sum}\to\infty$, и $h(PN)\to 0$, $NPh^n(NP)\to\infty$. Тогда

$$\tilde{p}_W(w/\omega)=\frac{1}{PNh^n}\sum_{s=1}^{N}\sum_{k=1}^{P}\varphi\left(\frac{w-(u^{(s)}+v^{(s,k)}+g^{(s,k)})}{h}\right)=$$
$$=\frac{1}{N}\sum_{s=1}^{N}\mu_P^s(w),$$

где каждая вложенная сумма $\mu_P^s(w)$ является ядерной оценкой и имеет при условии $u=u^{(s)}$ следующий вид:

$$\mu_P^s(w)=\frac{1}{Ph^n}\sum_{k=1}^{P}\varphi\left(\frac{w-(u^{(s)}+v^{(s,k)}+g^{(s,k)})}{h}\right)=$$
$$=\mu_P^s\left(w/u=u^{(s)}\right)=\frac{1}{Ph^n}\sum_{k=1}^{P}\varphi\left(\frac{v+g-(v^{(s,k)}+g^{(s,k)})}{h}\right).$$

При $P\to\infty$ выполняется следующее предельное соотношение:

$$\mu_P^s\left(w/u=u^{(s)}\right)\underset{P\to\infty}{\to}p_{W_s}(w/u^{(s)})=p_Q(w-u^{(s)})=$$
$$=\int p_G((w-u^{(s)})-v)p_V(v)\,\mathrm{d}v=$$
$$=\int\frac{1}{h'^n}\varphi\left(\frac{(w-u^{(s)})-v}{h'}\right)p_V(v)\,\mathrm{d}v=\vartheta_N\left(w-u^{(s)}\right).$$

Таким образом, полученная сумма $\tilde{p}_W(w/\omega)$ совпадает по форме с СЯО $\tilde{\tilde{p}}_X(x/\omega)$ плотности $p_X(x/\omega)$, а, значит, сгенерированные на ее основе данные могут быть использованы для обучения так же, как и данные, сгенерированные на основе $\tilde{p}_X(x/\omega)$ в соответствии с алгоритмом [4].

Очевидным следствием доказанного утверждения является то, что при достаточно больших $N\to\infty$ искусственное размножение данных может осуществляться как $w^{(s,k)}=u^{(s)}+v^{(s,k)}$, $k=\overline{1,P}$, так как $h'=h'(N)$ при этом стремится к нулю, а $\psi_N(g)$ стремится к дельта-функции.

Для смешанных оценок вида (3), (4) необходимо рассмотреть правила выбора оптимальных значений $\tilde{\tilde{h}}=h(U_{le}^{N})$ на основе анализа свойств обучающих выборок по каждому из классов. Известная методика [1] определения оптимального значения $\tilde{h}=h(U_{le}^{N})$ для обычной (не смешанной) оценки основана на нахождении максимума логарифма отношения правдоподобия при кросс-валидации обучающей выборки. Рассмотрим сначала обычную оценку (2)

$$\tilde{p}_U(u/\omega)=\frac{1}{Nh^n}\sum_{s=1}^{N}\varphi\left(\frac{u-u^{(s)}}{h}\right)=$$
$$=\frac{1}{N}\sum_{s=1}^{N}\psi_N\left(u-u^{(s)}\right).$$

Для нее задача максимизации логарифма отношения правдоподобия при кросс-валидации формулируется следующим образом:

$$\tilde{L}(h)=\ln\prod_{s=1}^{N}\tilde{p}_{U_s}(u^{(s)}/\omega)=$$
$$=\sum_{s=1}^{N}\ln\left(\sum_{p=1,p\neq s}^{N}\frac{1}{(N-1)h^n}\varphi\left(\frac{u^{(s)}-u^{(p)}}{h}\right)\right)\to\max_h,\qquad(7)$$

где $\tilde{p}_{U_s}(u^{(s)}/\omega)$ – ядерная оценка ПРВ, полученная по всем элементам обучающей выборки, за исключением $u^{(s)}$. В [1] также показано, что при использовании ядра с диагональной матрицей $\Xi=I$ область расположения максимума функции $\tilde{L}(h)$ для поиска оптимального значения путем перебора находится в определенных границах.

Сформулируем критерий нахождения оптимального $\tilde{\tilde{h}}=h(U_{le}^{N})$ для смешанной ядерной оценки. Сложность здесь состоит в том, что нужно провести кросс-валидацию, основываясь только на наблюдениях $U_{le}{}^{N}=\{u^{(1)},...,u^{(N)}\}$ при отсутствии изначально

наблюдений $X_{le}^N = \{x^{(1)}, ..., x^{(N)}\}$ величины, для которой строится плотность распределения. Введем случайные векторы $X^{(s)} = u^{(s)} + V^{(s)}$, $s = \overline{1, N}$, т.е. векторы, получаемые при независимом добавлении к каждому значению $u^{(s)}$ случайных векторов, имеющих известное распределение $p_V(v)$. Предлагается использовать следующий критерий для выбора параметра $\tilde{\tilde{h}} = h(U_{le}^N)$

$$\begin{aligned}
&\tilde{L}(h) = \ln \prod_{s=1}^{N} \int \tilde{\tilde{p}}_{X_s}(u^{(s)} + v^{(s)} / \omega) = \\
&= \sum_{s=1}^{N} \ln \left(\frac{1}{(N-1)} \times \right. \\
&\left. \times \sum_{\substack{p=1, \\ p \neq s}}^{N} \int \vartheta_N \left(u^{(s)} + v^{(s)} - u^{(p)} \right) p_V(v^{(s)}) \mathrm{d}v^{(s)} \right) \rightarrow \\
&\rightarrow \max_h, \\
&\vartheta_N \left(x - u^{(p)} \right) = \int \frac{1}{h^n} \varphi \left(\frac{x - v - u^{(p)}}{h} \right) p_V(v) \mathrm{d}v = \\
&= \int \psi_N \left(x - u^{(p)} - v \right) p_V(v) \mathrm{d}v.
\end{aligned} \tag{8}$$

Данный критерий определяет в качестве оптимального значения $\tilde{\tilde{h}} = h(U_{le}^N)$ такое значение, которое максимизирует математическое ожидание логарифма отношения правдоподобия, полученное при усреднении по всем возможным значениям случайных векторов $X^{(s)} = u^{(s)} + V^{(s)}$, $s = \overline{1, N}$. Учитывая избранный нами способ получения смешанной оценки, выражение (8) можно переписать в виде

$$\begin{aligned}
&\tilde{L}(h) = \ln \prod_{s=1}^{N} \int \tilde{\tilde{p}}_{X_s}(u^{(s)} + v^{(s)} / \omega) = \\
&= \sum_{s=1}^{N} \ln \left(\frac{1}{(N-1)} \times \right. \\
&\left. \times \sum_{\substack{p=1, \\ p \neq s}}^{N} \int \int \left(\frac{1}{h^n} \varphi \left(\frac{u^{(s)} + v^{(s)} - v^{(p)} - u^{(p)}}{h} \right) \times \right. \right. \\
&\left. \left. \times p_V(v^{(p)}) p_V(v^{(s)}) \right) \mathrm{d}v^{(p)} \mathrm{d}v^{(s)} \right) \rightarrow \max_h.
\end{aligned} \tag{9}$$

Очевидно, что для любой пары независимых значений $v^{(s)}$, $v^{(p)}$ случайных величин $V^{(s)}$, $V^{(p)}$ максимум интеграла в (9) относительно h достигается для максимума подынтегрального выражения.

Использование (9) с многомерным интегралом для нахождения оптимального значения $\tilde{\tilde{h}} = h(U_{le}^N)$ путём перебора в заданном диапазоне значений, за исключением случая гауссовской ПРВ $p_V(v)$, затруднено. Поэтому, опираясь на (9), можно рассмотреть следующий способ определения $\tilde{\tilde{h}} = h(U_{le}^N)$ с использованием СЯО, а именно: осуществлять максимизацию выборочного математического ожидания $\tilde{L}(h)$, построенного путём генерации для каждой точки $u^{(s)}$, $s = \overline{1, N}$ подвыборки $V_{le}^Q = \{v^{(s,1)}, ..., v^{(s,Q)}\}$, позволяющей сформировать значения $x^{(s,r)} = u^{(s)} + v^{(s,r)}$, $r = \overline{1, Q}$ для подстановки в (9). Тогда, заменяя интеграл суммой, величину, максимизируемую при кроссвалидации, можно приближённо представить следующим образом:

$$\begin{aligned}
&\tilde{L}(h) = \ln \prod_{s=1}^{N} \int \tilde{\tilde{p}}_{X_s}(u^{(s)} + v^{(s)} / \omega) \approx \\
&\approx \sum_{s=1}^{N} \ln \left(\sum_{\substack{p=1, \\ p \neq s}}^{N} \frac{1}{(N-1)Q} \sum_{r=1}^{Q} \vartheta_N \left(u^{(s)} + v^{(s,r)} - u^{(p)} \right) \right) \rightarrow \\
&\rightarrow \max_h.
\end{aligned} \tag{10}$$

Непосредственное использование (10) затратно в вычислительном отношении, поэтому целесообразно использовать итерационный алгоритм наращивания размера подвыборок $V_{le}^Q = \{v^{(s,1)}, ..., v^{(s,Q)}\}$ с запоминанием промежуточных результатов и контролем получаемых изменений $\tilde{\tilde{h}} = h(U_{le}^N)$ на соседних шагах изменения $Q = 1, 2, ...$

2. Результаты экспериментальных исследований и их обсуждение

2.1. Результаты моделирования к анализу свойств смешанных оценок

Для проведения экспериментальных исследований свойств смешанных оценок необходимо использовать численный показатель для сравнения ПРВ и оценки качества восстановления. В качестве такого показателя была выбрана широко используемая в аналогичных задачах интегральная квадратическая ошибка (ISE) [17–19]:

$$\text{ISE } \tilde{p} = \int \left(\tilde{p}(x) - p(x) \right)^2 \mathrm{d}x.$$

Данный показатель позволяет учитывать значительные и продолжительные отклонения сравниваемых ПРВ; при этом малые отклонения слабо сказываются на его величине. Кроме того, при небольших размерностях данных значение ISE $\tilde{p}$ может быть получено с помощью численных методов для функций $\tilde{p}(x)$ и $p(x)$ произвольной формы, что обеспечивает универсальность его применения в ходе экспериментов.

При экспериментальной проверке сходимости смешанной оценки $\tilde{\tilde{p}}(x / \omega)$ было проведено несколько экспериментов с использованием искусственно сгенерированных одномерных и двумерных данных. Для каждого типа ПРВ тестирование проводилось при трёх наборах параметров распределений, подобранных таким образом, чтобы были реализованы три варианта соотношения дисперсий информативной и помеховой составляющей $\rho = D[u]/D[v]$: $\rho > 1$, $\rho = 1$, $\rho < 1$. При этом для одномерных данных было рассмотрено три случая:

1. Распределение U – смесь из двух гауссовских распределений с математическими ожиданиями $M[u_1]=-2$, $M[u_2]=-1$ и дисперсиями $D[u_1]=1$, $D[u_2]=2$ ($\rho=1$, $\rho=4$), $D[u_1]=0{,}5$, $D[u_2]=1$ ($\rho=0{,}25$); распределение V – гауссовское с нулевым математическим ожиданием и дисперсией $D[v]=1$ ($\rho=4$), $D[v]=2$ ($\rho=0{,}25$, $\rho=1$).
2. Распределение U – равномерное в диапазоне [0...4] для $\rho=1$, $\rho=4$ и [0...2] для $\rho=0{,}25$; распределение V – равномерное с нулевым математическим ожиданием в диапазоне [–1...1] для $\rho=4$ и [–2...2] для $\rho=0{,}25$, $\rho=1$.
3. Распределение U – экспоненциальное с параметром $\lambda=\sqrt{1/2}$ ($\rho=1$, $\rho=4$), $\lambda=\sqrt{2}$ ($\rho=0{,}25$); распределение V – треугольное с нулевым математическим ожиданием в диапазоне $[-\sqrt{3}...\sqrt{3}]$ для $\rho=4$ и $[-2\sqrt{3}...2\sqrt{3}]$ для $\rho=0{,}25$, $\rho=1$.

Для двумерных данных рассматривалась мультигауссовская ПРВ, т.е. U формировалась на основе смеси двух гауссовских распределений с математическими ожиданиями $M[u_1]=(-1,-1)$, $M[u_1]=(1,2)$, а V – гауссовское распределение с нулевым математическим ожиданием. Матрицы ковариаций для всех распределений выбирались диагональными, причем элементы на главной диагонали совпадали с соответствующими дисперсиями для одномерного случая.

В ходе исследования проверялась скорость сходимости СЯО $\tilde{\tilde{p}}(x/\omega)$ к $p(x/\omega)$ при $N\to\infty$. Кроме того, проводилось сравнение смешанной оценки $\tilde{\tilde{p}}(x/\omega)$ с обычной ядерной оценкой $\tilde{p}(x/\omega)$. Для этого в начале каждого эксперимента генерировалось N значений случайной величины U, по которым методом Парзена восстанавливалась плотность распределения $\tilde{p}_U(u/\omega)$, а затем вычислялась смешанная оценка $\tilde{\tilde{p}}(x/\omega)$. Также с использованием полученных значений U генерировалось N значений случайной величины $X=U+V$, по которым вычислялась обычная оценка Парзена $\tilde{p}_X(x/\omega)$. Исследования проводились для значений N от 50 до 5000, при этом для каждого N выполнялось по 100 экспериментов (для двумерного случая – по 10 экспериментов), в которых вычислялись усредненные значения ISE для оценок $\tilde{p}(x/\omega)$ и $\tilde{\tilde{p}}(x/\omega)$. Полученные зависимости интегральной квадратичной ошибки оценок плотностей $\tilde{p}(x/\omega)$ и $\tilde{\tilde{p}}(x/\omega)$ от объёма исходных данных приведены на рис. 1.

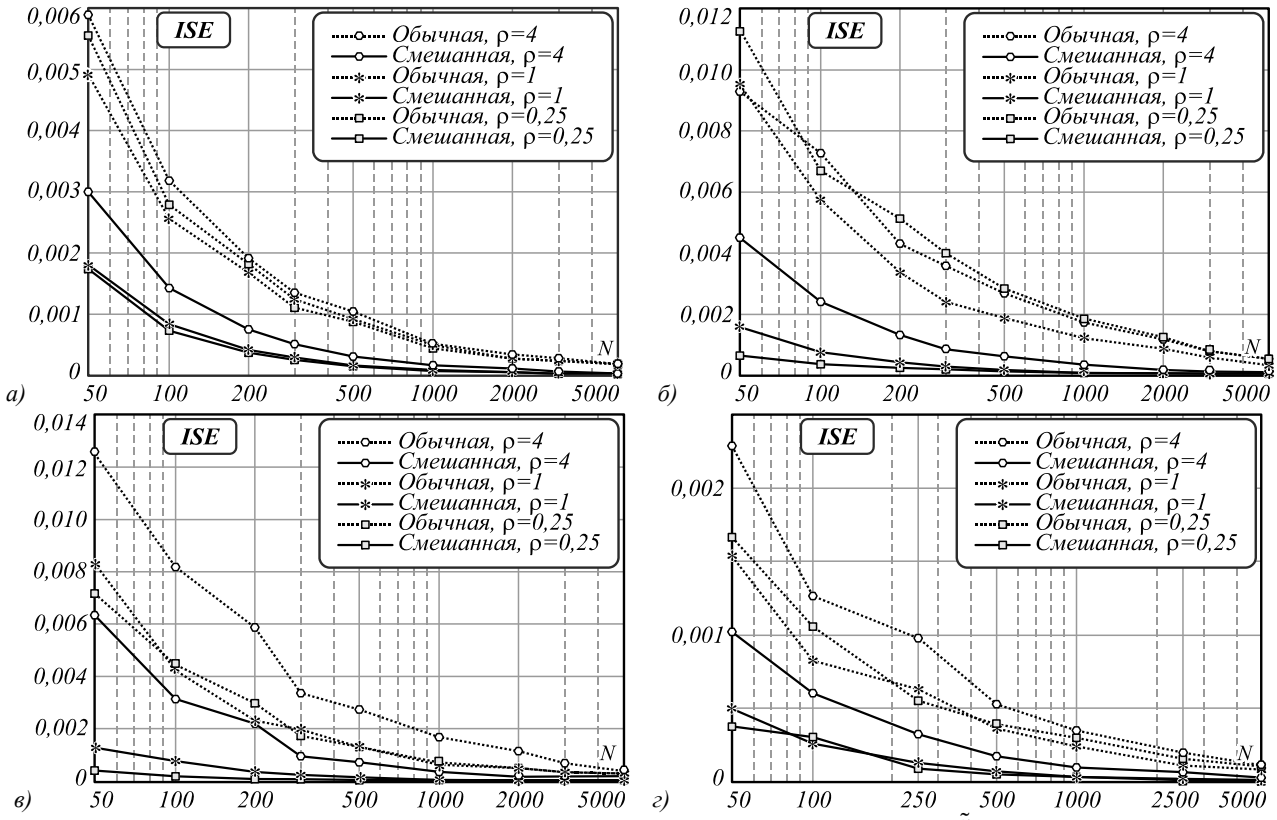

Рис. 1. Зависимости интегральной квадратичной ошибки оценок плотностей $\tilde{p}(x/\omega)$ и $\tilde{\tilde{p}}(x/\omega)$ от объёма исходных данных для одномерного мультигауссовского распределения (а), одномерного равномерного распределения (б), одномерных экспоненциального и треугольного распределений (в) и двумерного мультигауссовского распределения (г)

Из графиков видно, что для всех рассмотренных случаев при увеличении объёма данных N среднеквадратичная ошибка монотонно убывает как для $\tilde{p}(x/\omega)$, так и для $\tilde{\tilde{p}}(x/\omega)$. При этом величина ошибки для смешанной оценки $\tilde{\tilde{p}}(x/\omega)$ оказывается значительно меньше при любых N, что подтверждается и визуальным сравнением формы получаемых ПРВ, один из примеров которых приведён на рис. 2.

На рис. 1 также видно, что для всех рассматриваемых распределений значения ошибки для СЯО при $\rho = 1$ и $\rho = 0{,}25$ ниже, чем при $\rho = 4$, и практически совпадают между собой. Для обычных оценок аналогичная ситуация наблюдается только для случая экспоненциального распределения, тогда как для случая одномерного мультигауссовского распределения величина ошибки уже практически не зависит от ρ при любом размере выборки. Для равномерного распределения минимальная ошибка при использовании обычных оценок Парзена достигается при $\rho = 1$, максимальная – при $\rho = 0{,}25$, для двумерного мультигауссовского – при $\rho = 1$ и $\rho = 4$ соответственно.

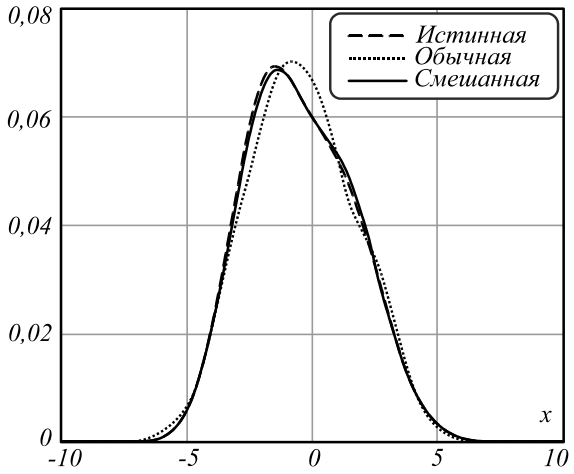

Рис. 2. Пример восстановленных функций плотности для случая одномерного мультигауссовского распределения при $N = 300$ и $\rho = 4$

Заметный рост величины ошибки для СЯО при значениях дисперсии величины U, превышающих дисперсию величины V, по-видимому, обусловлен тем, что точность смешанных оценок полностью зависит от точности оценки плотности U, а большие значения дисперсии при том же объёме обучающей выборки приводят к снижению точности оценки.

2.2. Исследование применимости смешанных ядерных оценок для аугментации данных

При решении данной задачи был проведен эксперимент, направленный на сравнение эффективности применения СЯО в задачах классификации образов на основе искусственных нейронных сетей (ИНС), а также других типов классификаторов. В ходе эксперимента сравнивалась точность классификации многомерных случайных величин нескольких классов ω_i, описываемых векторами вида $X_i = U_i + V$, $i = \overline{1, M}$. Рассматривались следующие типы классификаторов:

– ИНС, обученная по малой обучающей выборке, содержащей N образцов каждого класса;
– ИНС, обученная по большой обучающей выборке, содержащей N' образцов каждого класса;
– ИНС, обученная по обучающей выборке, искусственно размноженной из N до N' образцов в соответствии с результатами утверждения 3;
– байесовский классификатор, синтезированный для плотности распределения данных $\tilde{\tilde{p}}(x/\omega_i)$, полученной на основе смешанной ядерной оценки;
– байесовский классификатор, синтезированный для известной плотности распределения данных $p(x/\omega_i)$ и используемый для оценки нижней границы вероятности ошибок.

В качестве U_i использовались случайные векторы, имеющие для каждого класса ω_i ПРВ мультигауссовского вида, т.е. взвешенной с весами $p^{(i)} = \{p_1^{(i)}, ..., p_{m_i}^{(i)}\}$ суммы m_i компонент:

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

$$b_s^{(i)}(u) = N\left(u, M[u_s^{(i)}], C_s^{(i)}\right), \sum_{s=1}^{m_i} p_s^{(i)} = 1, i = \overline{1, M},$$

где $p_s^{(i)}$ – вероятность появления вектора с параметрами, соответствующими данной компоненте смеси.

В ходе эксперимента генерировались значения случайных векторов, соответствующие трем классам образов, представленных смесями с несколькими компонентами внутри каждого класса (по две компоненты для первого и второго классов и три компоненты для третьего) со следующими параметрами:

$$p^{(1)} = \{0{,}5, 0{,}5\}, p^{(2)} = \{0{,}5, 0{,}5\},$$
$$p^{(3)} = \{0{,}33, 0{,}33, 0{,}34\},$$
$$M\left[u_s^{(i)}\right] = \Delta^{(i)} + \mathrm{rand}(0, \mathrm{d}\mu),$$
$$\Delta^{(1)} = 0, \Delta^{(2)} = 1, \Delta^{(3)} = 2,$$
$$C_s^{(i)} = \left\| c_{kt}^{(s,i)} \right\|, \; c_{kt}^{(s,i)} = \left(r_s^{(i)}\right)^{|k-t|},$$

где $\mathrm{rand}(0, \mathrm{d}\mu)$ – равномерно распределенная случайная величина в указанном диапазоне; $\mathrm{d}\mu$ – 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 $[r_{\min}, r_{\max}]$. Для V использовалось гауссовское распределением с нулевым математическим ожиданием и единичной матрицей ковариации $C_v = I$. Примеры формирования малой выборки, иллюстрирующие влияние параметров $\mathrm{d}\mu$ и r на генерируемые данные, показаны на рис. 3.

В эксперименте для каждого класса генерировалось $N = 100$, $N' = 200..1000$ и $\tilde{N} = 1000$ значений случайной величины $X_i = U_i + V$, составляющих соответственно малую обучающую выборку, большую обучающую выборку и тестовую выборку. Эти выборки использовались для обучения однотипных классификаторов: на основе малой выборки, на основе выборки, полученной путём искусственного размножения малой выборки до N' образцов, а также на основе большой выборки, объём которой устанавливался равным искусственно размноженной выборке. Объём тестовых выборок во всех случаях составлял

$\tilde{N} = 1000$ образов для каждого класса. Каждый образ описывается $n = 20$ признаками, количество которых сокращалось при обучении нейронных сетей до 5 с помощью метода главных компонент [20]. Для каждого значения N' выполнялось по 1000 экспериментов, после чего полученные значения усреднялись.

Используемая в качестве обучаемого классификатора нейронная сеть является сетью прямого распространения класса MLP (многослойный персептрон) [21]. Сеть содержит один скрытый слой с сигмоидальной функцией активации [22] и один выходной слой с линейной функцией активации [23]. Количество входных контактов сети соответствует количеству используемых признаков распознавания, а количество нейронов в выходном слое m_2 равно числу классов. 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 2013b, для обучения использовался алгоритм Левенберга–Марквардта [24]. В табл. 1 приведены результаты, полученные для смесей гауссовских случайных векторов со слабо коррелированными признаками ($r_{min} = 0{,}4$, $r_{max} = 0{,}5$), а в табл. 2 – с сильно коррелированными признаками ($r_{min} = 0{,}8$, $r_{max} = 0{,}9$).

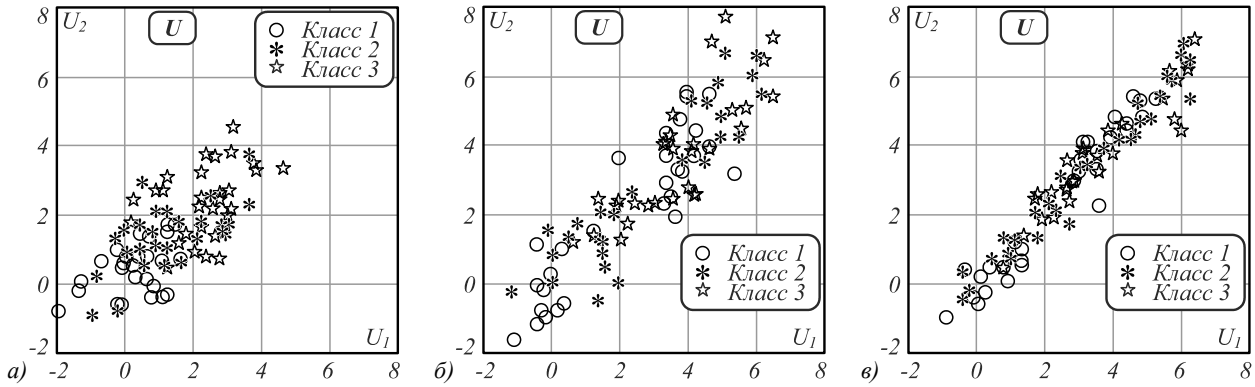

Рис. 3. Примеры формирования малой выборки:
$d\mu = 0{,}5$, $r_{min} = 0{,}4$, $r_{max} = 0{,}5$ (а), $d\mu = 4$, $r_{min} = 0{,}4$, $r_{max} = 0{,}5$ (б), $d\mu = 4$, $r_{min} = 0{,}8$, $r_{max} = 0{,}9$ (в)

Табл. 1. Вероятности ошибочного распознавания для случая слабо коррелированных данных

dμ	Классификатор	Вероятность ошибочн. распознавания N'				
		200	400	600	800	1000
0,5	*HC* (малая выборка)	0,1833	0,1828	0,1834	0,1832	0,1844
	HC (большая выб.)	0,1672	0,1603	0,1587	0,1572	0,1570
	HC (размн. добавл. *V*)	0,1756	0,1664	0,1635	0,1625	0,1634
	Байес. класс. (смеш.)	0,1864	0,1867	0,1867	0,1865	0,1868
	Байесовский классиф.	0,1479	0,1481	0,1480	0,1484	0,1483
2	*HC* (малая выборка)	0,1536	0,1553	0,1556	0,1552	0,1532
	HC (большая выб.)	0,1407	0,1346	0,1310	0,1303	0,1305
	HC (размн. добавл. *V*)	0,1510	0,1403	0,1382	0,1373	0,1372
	Байес. класс. (смеш.)	0,1100	0,1105	0,1105	0,1098	0,1105
	Байесовский классиф.	0,0771	0,0775	0,0775	0,0774	0,0775
4	*HC* (малая выборка)	0,0569	0,0592	0,0597	0,0576	0,0575
	HC (большая выб.)	0,0478	0,0407	0,0396	0,0365	0,0359
	HC (размн. добавл. *V*)	0,0539	0,0468	0,0454	0,0420	0,0413
	Байес. класс. (смеш.)	0,0157	0,0157	0,0163	0,0157	0,0157
	Байесовский классиф.	0,0087	0,0087	0,0090	0,0087	0,0087

Анализ результатов показывает, что независимо от степени корреляции и пересечения разных классов использование предлагаемого алгоритма ИРД позволяет значительно сократить вероятность ошибочного распознавания с помощью нейронной сети при наличии ограниченного объёма обучающих данных.

Байесовский классификатор, синтезированный на основе смешанных ядерных оценок ПРВ, обеспечивает высокую точность распознавания для умеренно и сильно рассредоточенных данных ($d\mu \ge 2$), значительно превосходя нейронные сети, обученные по большой и размноженным выборкам. Для сильно смешанных данных ($d\mu \ge 0{,}5$) его точность становится сопоставимой с нейронной сетью, обученной по малой выборке. При этом все алгоритмы демонстрируют похожие результаты с высоким процентом ошибочных распознаваний (32 – 35 %), а минимально достижимый процент ошибок при использовании байесовского классификатора с известной плотностью распределения составляет всего 29 %. Это обусловлено малой удалённостью в многомерном пространстве значений U разных классов друг от друга относительно уровня аддитивного шума V, воздействие которого приводит к существенному пересечению данных разных классов, что нивелирует эффект от применения смешанной оценки.

2.3. Экспериментальные исследования алгоритмов распознавания биологических объектов на основе спектральных измерений

Проверка возможности использования смешанных оценок для ИРД с целью повышения качества классификатора при малых обучающих выборках выполнялась в задаче классификации здоровых и поражен-

ных грибковыми заболеваниями элементов зерновых смесей (ЭЗС) (на примере ЭЗС ячменя, пораженных фузариозом). Разработанная для этих целей установка экспресс-анализа и классификации элементов неоднородного потока зерновых смесей [25] представлена на рис. 4 и позволяет анализировать спектры отражения и/или пропускания в видимом и ближнем ИК-диапазоне.

Табл. 2. Вероятности ошибочного распознавания для случая сильно коррелированных данных

dμ	Классификатор	Вероятность ошибочн. распознавания N'				
		200	400	600	800	1000
0,5	*HC* (малая выборка)	0,3511	0,3517	0,3495	0,3513	0,3495
	HC (большая выб.)	0,3332	0,3209	0,3157	0,3139	0,3116
	HC (размн. добавл. *V*)	0,3377	0,3266	0,3242	0,3233	0,3215
	Байес. классиф. (смеш.)	0,3412	0,3419	0,3422	0,3418	0,3409
	Байесовский классиф.	0,2886	0,2893	0,2890	0,2890	0,2882
2	*HC* (малая выборка)	0,2650	0,2673	0,2635	0,2640	0,2642
	HC (большая выб.)	0,2440	0,2328	0,2299	0,2266	0,2241
	HC (размн. добавл. *V*)	0,2658	0,2555	0,2524	0,2491	0,2456
	Байес. классиф. (смеш.)	0,1534	0,1541	0,1541	0,1533	0,1524
	Байесовский классиф.	0,1075	0,1084	0,1085	0,1077	0,1072
4	*HC* (малая выборка)	0,1026	0,1014	0,1019	0,1002	0,1011
	HC (большая выб.)	0,0829	0,0722	0,0700	0,0674	0,0657
	HC (размн. добавл. *V*)	0,0996	0,0881	0,0854	0,0833	0,0810
	Байес. классиф. (смеш.)	0,0134	0,0133	0,0139	0,0137	0,0135
	Байесовский классиф.	0,0062	0,0062	0,0064	0,0065	0,0062

Установка основана на использовании оптоволоконной техники. В ней анализируемые ЭЗС из накопителя 1 с помощью устройства транспортировки 2, состоящего из вибропитателя и скатного профилированного лотка, распределяются по каналам, один из которых содержит зону спектрального анализа 3, в которой происходит регистрация спектров отражения (или люминесценции), и коллимированного пропускания (рассеяния вперед) в видимом и ближнем ИК-диапазоне (от 350 до 2 200 нм) с помощью волоконных спектрометров 9 и 12. Для регистрации спектров в диапазоне от 350 до 1000 нм использован спектрометр Ocean Optics USB4000-VIS-NIR [26], в диапазоне от 900 до 2200 нм – спектрометр Ocean Optics NirQuest512-2.2 [27]. Для освещения области спектрального анализа зерна 3 используется волоконный галогеновый источник излучения 4 Ocean Optics HL-2000 [28], сопряженный с волноводом 5 и объективом 6.

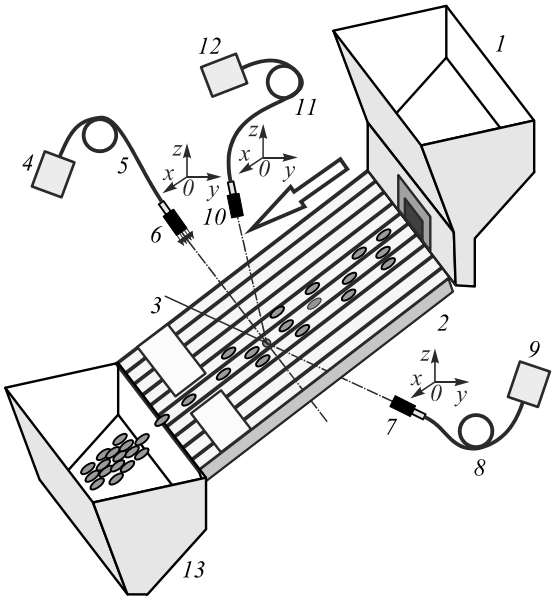

Рис. 4. Схема установки потокового экспресс-анализа ЭЗС

Прошедшее через ЭЗС излучение (рассеянное вперед излучение) собирается широкоапертурным объективом 7, сопряжённым с выходным концом световода 8, и передаётся в спектрометр 9, соединённый с компьютером. Отражённое от ЭЗС излучение собирается широкоапертурным объективом 10, сопряжённым с выходным концом световода 11, и передаётся в спектрометр 12, соединённый с компьютером. После прохождения зоны спектрального анализа 3 ЭЗС попадают в накопительный бункер 13. Подробное описание установки содержится в работе [25].

В качестве примера внесения в СЯО неинформативной помеховой составляющей V рассматривались флуктуации спектров пропускания, обусловленные вращением ЭЗС в процессе попадания и перемещения в зону спектрального анализа. Главным фактором, определяющим флюктуации формы спектров в данном случае, является неоднородность формы и распределения поражённых участков по поверхности ЭЗС.

При решении этой задачи на основе экспериментальных данных проводилось независимое построение оценки плотности $\tilde{p}_V(v)$, которая использовалась для получения смешанной оценки. С целью получения необходимых данных были проведены измерения спектров пропускания в видимом и ближнем ИК-диапазоне (от 350 до 1000 нм) для 10 здоровых и 10 поражённых ЭЗС с поворотом на углы с шагом в 2 градуса с помощью установки, представленной на рис. 5.

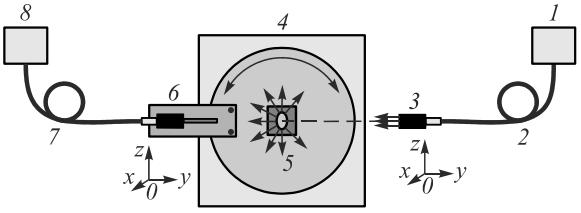

Рис. 5. Схема установки для анализа флуктуаций спектров при вращении ЭЗС

В этой установке ЭЗС 5 помещаются в центре вращающегося столика 4 (Thorlabs PR01/M4 [29]), имеющего сквозное отверстие. В результате при вращении столика ЭЗС остаётся неподвижным. Волоконный галогеновый источник излучения 1 HL-2000 [28], сопряжённый с волноводом 2 и объективом 3, формирует квазиколлимированный световой пучок, который направляется на образец 5. Числовая апертура волокна $NA = 0{,}22$, диаметр волновода – 600 мкм. Отметим, что диаметр светового пучка можно изменять как в меньшую сторону, так и в большую, в зависимости от конкретной задачи. Объектив 3 содержит асферическую линзу с числовой апертурой $NA = 0{,}51$. Диаметр падающего пучка света составляет величину около 1 мм, который существенно ниже средней ширины исследуемых зерен, составляющей величину около 3–4 мм. Прошедшее (рассеянное) через ЭЗС излучение собирается широкоапертурным объективом 6, сопряжённым с выходным концом световода 7 (числовая апертура волокна $NA = 0{,}5$, диаметр волновода 1000 мкм) и передаётся в спектрометр 8, соединённый с компьютером. Объектив 7 содержит асферическую линзу с числовой апертурой $NA = 0{,}51$. На объектив по направлению главной оптической оси устанавливается трубка для обеспечения возможности регистрации рассеянного от зерна излучения в пределах малого телесного угла. Расстояние между объективом 6 и зерном фиксировано и составляет величину около 40 мм. Объектив 6 устанавливается на пластину, которая крепится к поверхности вращающегося столика 4. В результате от каждого ЭЗС возможна регистрация спектров пропускания (рассеяния вперед) с шагом 2 градуса в видимом и ближнем ИК-диапазоне с помощью волоконного спектрометра 8 (USB4000-VIS-NIR [26] либо NirQuest512-2.2 [27]) в пределах от 0 до 180 градусов. Для юстировки объективов 3 и 6 используются специальные трёхкоординатные подвижки Thorlabs DT12XYZ/M [30].

Количество спектральных составляющих, измеряемых по каждому ЭЗС, определяющее количество исходных признаков классификации, составляло 2287.

В ходе обработки спектральных измерений по последовательности $X^{(s)} = \{x^{(1)},...,x^{(N_s)}\}$ для каждого из зёрен $s = 1..20$ вычислялось значение $u^{(s)} = M[X^{(s)}]$, которое рассматривалось как информативная составляющая, описывающая свойства непосредственно зерна, и значения

$$V^{(s)} = \{v^{(1)},...,v^{(N_s)}\} = \{x^{(1)} - u^{(s)},...,x^{(N_s)} - u^{(s)}\},$$

$$M\left[V^{(s)}\right] = 0,$$

соответствующие искажениям, возникающим при повороте данного зерна. После обработки данных по всем зёрнам из значений $V^{(s)}$, $s = 1..20$ формировалась итоговая выборка $V_{le}^{N_v}$, $N_v = 168$. Затем размерность данных сокращалась с исходных 2287 признаков до $n = 5$ с помощью метода главных компонент [20], и с помощью метода Парзена вычислялась n-мерная ядерная оценка плотности $\tilde{p}_V(v)$. Необходимость сокращения размерности исходного признакового пространства обусловлена известным общим свойством ухудшения свойств ядерных оценок при больших размерностях вектора признаков [1]. Выбор указанной размерности вектора признаков после преобразования по методу главных компонент обусловлен данными, полученными в ранее выполненных авторами работах (например, [25]), в которых показано, что понижение размерности данных спектральных измерений без существенной потери информативности целесообразно проводить до значений $n = 5...10$.

В качестве данных для обучения и тестирования алгоритма классификации использовались измеренные на той же установке в потоковом режиме (без целенаправленного поворота зёрен на различные углы) 546 спектров для здоровых зёрен ячменя и 543 для поражённых фузариозом.

Размножение данных в обучающей выборке с N до N' образцов для каждого класса выполнялось путём добавления к имеющимся N спектрам пропускания искусственно генерируемых на основе плотности $\tilde{p}_V(v)$. Размер размноженной выборки выбирался кратным исходному ($N' = N + kN$), где каждый из N исходных образцов попадал в обучающую выборку один раз в неизменном виде и k раз с добавлением случайной величины V. Для сравнения также выполнялось ИРД по методу [4], основанному на генерации данных с использованием традиционной оценки плотности $\tilde{p}_X$, получаемой по обучающей выборке.

В качестве классификатора для тестирования использовалась нейронная сеть класса MLP [21] с одним скрытым слоем с сигмоидальной функцией активации [22] и одним выходным слоем с линейной функцией активации [23]. Количество входных контактов сети соответствует количеству используемых признаков $n = 5$, количество нейронов в выходном слое m_2 равно числу классов (в данном случае 2).

При тестировании использовался метод скользящего контроля. Его суть состояла в изъятии из исходной выборки случайным образом 5 % ($N = 27$) образцов каждого класса, проведении размножения данных и обучения алгоритма на данных образцах и последующем тестировании с использованием оставшихся 95 % образцов ($\tilde{N} = 519$ для здоровых зёрен и $\tilde{N} = 516$ для заражённых). Полученные результаты усреднялись по 10000 итерациям. Тестирование проводилось для обучающих выборок объёмом от $N' = N = 27$ образцов каждого класса (не содержащих искусственно сгенерированных данных) до $N' = 6N = 162$ (содержащих $5N$ искусственно сгенерированных образцов каждого класса). Аналогичный эксперимент был проведён с использованием в качестве обучающей выборки 10 % ($N = 54$) образцов каждого класса и последующем тестированием с использованием оставшихся 90 % образцов. Полученные результаты приведены на рис. 6. Для исходной выборки

представленные данные характеризуют результаты обучения для различных случайно выбираемых наборов данных фиксированного объёма.

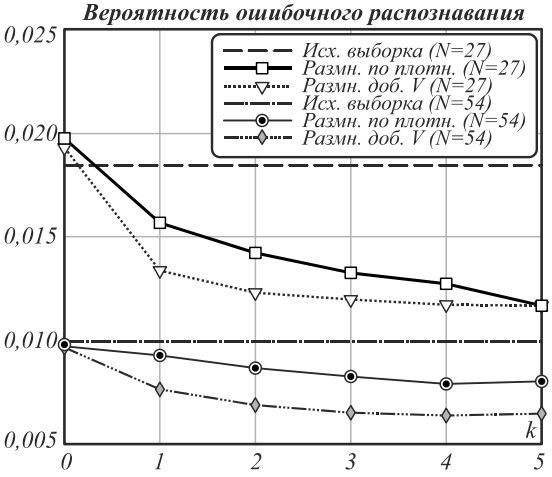

Рис. 6. Зависимости вероятности ошибочного распознавания зараженных семян ячменя от коэффициента k искусственного размножения данных в обучающей выборке при использовании выборок, содержащих N = 27 и N = 54 образцов каждого класса

Из графиков видно, что размножение данных в обучающей выборке на основе смешанных оценок путём добавления искусственно генерируемых величин V к спектрам из исходной обучающей выборки позволяет получить значительно меньший процент ошибочных распознаваний по сравнению с использованием обычной выборки. Так, для нейронной сети, обученной по малым выборкам с $N=27$ образцами каждого класса, частота ошибок составляет около 1,8 – 1,9 %, тогда как для нейронной сети, обученной по размноженной выборке, она меняется от 1,33 % (при $k=1$) до 1,16 % (при $k=5$). Алгоритм размножения данных с использованием традиционной оценки плотности $\tilde{p}_X$ в данной задаче оказался менее эффективным, однако позволил уменьшить частоту ошибок до 1,57 % при $k=1$ и 1,17 % при $k=5$. Схожая ситуация наблюдается и при использовании более крупных выборок ($N=54$).

Заключение

В работе предложен и исследован метод восстановления функции плотности распределения данных с использованием смешанных непараметрических ядерных оценок функций правдоподобия классов. В процессе исследования проведено сравнение предложенного метода с традиционными ядерными оценками, применяемыми напрямую к исходной обучающей выборке, показана сходимость смешанных оценок при увеличении объёма обучающей выборки, теоретически и экспериментально подтверждено, что использование алгоритмов, основанных на смешанных ядерных оценках, эквивалентно реализации процедуры искусственного размножения обучающих данных. Анализ полученных результатов позволяет сделать вывод о том, что в задачах восстановления ПРВ предлагаемый метод позволяет приблизительно вдвое снизить величину интегральной квадратичной ошибки для восстановленной плотности независимо от формы исходного распределения и размера обучающей выборки. На примере задачи классификации по спектрам пропускания в видимом и ближнем ИК-диапазоне здоровых и поражённых грибковыми заболеваниями элементов зерновых смесей была показана целесообразность использования смешанных ядерных оценок и базирующихся на их основе алгоритмов искусственного размножения данных.

Благодарности

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».

Литература

1. **Кривенко, М.П.** Непараметрическое оценивание элементов байесовского классификатора / М.П. Кривенко // Информатика и её применения. – 2010. – Т. 4, № 2. – С. 13-24.
2. **Лапко, А.В.** Непараметрический алгоритм автоматической классификации в условиях статистических данных большого объема / А.В. Лапко, В.А. Лапко // Информатика и системы управления. – 2018. – № 3(57). – С. 59-70. – DOI: 10.22250/isu.2018.57.59-70.
3. **Nakamura, Y.** Nonparametric density estimation based on self-organizing incremental neural network for large noisy data / Y. Nakamura, O. Hasegawa // IEEE Transactions on Neural Networks and Learning Systems. – 2016. – Vol. 28, Issue 1. – P. 8-17. – DOI: 10.1109/TNNLS.2015.2489225.
4. **Донских, А.О.** Метод искусственного размножения данных в задачах машинного обучения с использованием непараметрических ядерных оценок плотности распределения вероятностей / А.О. Донских, А.А. Сирота // Вестник Воронежского государственного университета Серия: Системный анализ и информационные технологии. – 2017. – № 3. – С. 142-155.
5. **Yaeger, L.** Effective training of a neural network character classifier for word recognition / L. Yaeger, R. Lyon, B. Webb // Advances in Neural Information Processing Systems 9 (NIPS 1996). – 1996. – P. 807-813.
6. **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. – Vol. 22, Issue 12 – P. 3207-3220. – DOI: 10.1162/NECO_a_00052.
7. **Simard, P.Y.** Best practices for convolutional neural networks applied to visual document analysis / P.Y. Simard, D. Steinkraus, J.C. Platt // Seventh International Conference on Document Analysis and Recognition. – 2003. – P. 958-963. – DOI: 10.1109/ICDAR.2003.1227801.
8. **Качалин, С.В.** Повышение устойчивости обучения больших нейронных сетей дополнением малых обучающих выборок примеров-родителей, синтезированными биометрическими примерами-потомками / С.В. Качалин // Труды научно-технической конференции кластера пензенских предприятий, обеспечивающих безопасность информационных технологий. – 2014. – Т. 9. – С. 32-35.
9. **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. – DOI: 10.18287/2412-6179-2016-40-6-911-918.
10. **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. – Vol. 6, Issue 1. – P. 30-39. – DOI: 10.1145/1007730.1007736.
11. **Chawla, N.V.** SMOTE: synthetic minority over-sampling technique / N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer // Journal of Artificial Intelligence Research. – 2002. – Vol. 16, Issue 1. – P. 321-357. – DOI: 10.1613/jair.953.
12. **Chawla, N.V.** SMOTEBoost: Improving prediction of the minority class in boosting / N.V. Chawla, A. Lazarevic, L.O. Hall, K.W. Bowyer. – In: Knowledge discovery in databases / ed. by N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel. – 2003. – P. 107-119.– DOI: 10.1007/978-3-540-39804-2_12.
13. **Фукунага, К.** Введение в статистическую теорию распознавания образов / К. Фукунага. – М.: Наука, 1979. – 368 с.
14. **Duda, R.O.** Pattern classification / R.O. Duda, P.E. Hart, D.G. Stork. – 2nd ed. – Hoboken, NJ: Wiley-Interscience, 2000. – 680 p.
15. **Крянев, А.В.**Математические методы обработки неопределенных данных / А.В. Крянев, Г.В. Лукин. – М.: Физмалит, 2003. – 216 с.
16. **Акимов, А.В.** Модели и алгоритмы распознавания цифровых изображений в условиях воздействия деформирующих и аддитивных искажений / А.В. Акимов, А.О. Донских, А.А. Сирота // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. – 2018. – № 1. – С. 104-118.
17. **Gramacki, A.** Nonparametric kernel density estimation and its computational aspects / A. Gramacki. – Cham, Switzerland: Springer International Publishing AG, 2018. – P. 42-49. – ISBN: 978-3-319-71687-9.
18. **Добровидов, А.В.** Выбор ширины окна ядерной функции в непараметрической оценке производной плотности методом сглаженной кроссвалидации / А.В. Добровидов, И.М. Рудько // Автоматика и телемеханика. – 2010. – № 2 – С. 42-58.
19. **Воронов, И.В.** Выбор ширины окна при аппроксимации плотности распределения вероятности методом Парзена-Розенблатта в случае малого объема выборки / И.В. Воронов, Р.Н. Мухометзянов, А.А. Краснова // Радиоэлектронная техника. – 2016. – № 1(9) – С. 93-98.
20. **Donskikh, A.O.** Optical methods of identifying the varieties of the components of grain mixtures based on using artificial neural networks for data analysis / A.O. Donskikh, D.A. Minakov, A.A. Sirota // Journal of Theoretical and Applied Information Technology – 2018. – Vol. 96, Issue 2. – P. 534-542.

Сведения об авторах

Сирота Александр Анатольевич, 1954 года рождения, в 1976 году окончил Воронежский государственный университет по специальности «Радиофизика и электроника». Доктор технических наук (1995 год), профессор, заведует кафедрой технологий обработки и защиты информации Воронежского государственного университета. Область научных интересов: синтез и анализ систем сбора и обработки информации, методы и технологии компьютерного моделирования информационных процессов и систем, системный анализ в сфере информационной безопасности, компьютерная обработка изображений, нейронные сети и нейросетевые технологии в системах принятия решений. E-mail: *sir@cs.vsu.ru* .

Донских Артём Олегович, 1993 года рождения, в 2016 году окончил магистратуру Воронежского государственного университета по специальности «Информационные системы и технологии». Аспирант кафедры технологий обработки и защиты информации Воронежского государственного университета. Область научных интересов: машинное обучение, компьютерная обработка изображений, искусственное размножение данных.
E-mail: *a.donskikh@outlook.com* .

Акимов Алексей Викторович, 1990 года рождения, в 2013 году окончил магистратуру Воронежского государственного университета по специальности «Информационные системы и технологии». Аспирант кафедры технологий обработки и защиты информации Воронежского государственного университета. Область научных интересов: распознавание изображений, машинное обучение. E-mail: *akimov@rcnit.vsu.ru* .

Минаков Дмитрий Анатольевич, 1982 года рождения, в 2005 году окончил Воронежский государственный университет по специальности «Оптика». Кандидат физико-математических наук (2008 год), старший научный сотрудник физической лаборатории факультета компьютерных наук Воронежского государственного университета. Область научных интересов: машинное обучение, спектральные методы анализа, оптоволоконная техника. E-mail: *minakov_d_a@mail.ru* .

ГРНТИ: 27.43.51, 28.23.37, 59.45.37
Поступила в редакцию 15 марта 2019 г. Окончательный вариант – 10 апреля 2019 г.

Multivariate mixed kernel density estimators and their application in machine learning for classification of biological objects based on spectral measurements

A.A. Sirota [1], A.O. Donskikh [1], A.V. Akimov [1], D.A. Minakov [1]
[1] Voronezh State University, Voronezh, Russia

Abstract

A problem of non-parametric multivariate density estimation for machine learning and data augmentation is considered. A new mixed density estimation method based on calculating the convolution of independently obtained kernel density estimates for unknown distributions of informative features and a known (or independently estimated) density for non-informative interference occurring during measurements is proposed. Properties of the mixed density estimates obtained using this method are analyzed. The method is compared with a conventional Parzen-Rosenblatt window method applied directly to the training data. The equivalence of the mixed kernel density estimator and the data augmentation procedure based on the known (or estimated) statistical model of interference is theoretically and experimentally proven. The applicability of the mixed density estimators for training of machine learning algorithms for the classification of biological objects (elements of grain mixtures) based on spectral measurements in the visible and near-infrared regions is evaluated.

Keywords: machine learning, pattern classification, data augmentation, kernel density estimation, spectral measurements.

Citation: Sirota AA, Donskikh AO, Akimov AV, Minakov DA. Multivariate mixed kernel density estimators and their application in machine learning for classification of biological objects based on spectral measurements. Computer Optics 2019; 43(4): 677-691. DOI: 10.18287/2412-6179-2019-43-4-677-691.

Acknowledgements: The presented study was supported by the Ministry of Education and Science of the Russian Federation under project No. 8.3844.2017/4.6 ("Development of facilities for express analysis and classification of the components of nonuniform grain mixtures with pathologies based on the integration between spectral analysis methods and machine learning")

References

[1] Krivenko MP. Nonparametric estimation of Bayesian classifier elements [In Russian]. Informatics and Applications 2010; 4(2): 13-24.

[2] Lapko AV, Lapko VA. Nonparametric algorithm of automatic classification under conditions of large-scale statistical data [In Russian]. Information Science and Control Systems 2018; 3(57): 59-70. DOI: 10.22250/isu.2018.57.59-70.

[3] Nakamura Y, Hasegawa O. Nonparametric density estimation based on self-organizing incremental neural network for large noisy data. IEEE Transactions on Neural Networks and Learning Systems 2016; 28(1): 8-17. DOI: 10.1109/TNNLS.2015.2489225.

[4] Donskikh AO, Sirota AA. A data augmentation method for machine learning based on nonparametric kernel density estimation [In Russian]. Proceedings of Voronezh State University. Series: system analysis and information technology 2017; 3: 142-155.

[5] Yaeger L, Lyon R, Webb B. Effective training of a neural network character classifier for word. NIPS 1996: 807-813.

[6] Ciresan DC, Meier U, Gambardella LM, Schmidhuber J. Deep big simple neural nets excel on handwritten digit recognition. Neural Computation 2010; 22(12): 3207-3220. DOI: 10.1162/NECO_a_00052.

[7] Simard PY, Steinkraus D, Platt JC. Best practices for convolutional neural networks applied to visual document analysis. 7th Int Conf Docum Anal Recogn 2003: 958-963. DOI: 10.1109/ICDAR.2003.1227801.

[8] Kachalin SV. Improving the stability of large neural networks by extending small training sets of parent samples with synthesized biometric descendant samples [In Russian]. Proceedings of the Scientific and Technical Conference of Thecluster of Penza Enterprises Providing Security of Information Technologies 2014; 9: 32-35.

[9] Akimov AV, Sirota AA. Synthetic data generation models and algorithms for training image recognition algorithms using the Viola-Jones framework. Computer Optics 2016; 40(6): 911-918. DOI: 10.18287/2412-6179-2016-40-6-911-918.

[10] Guo H, Viktor HL. Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach. ACM SIGKDD Explorations Newsletter 2004; 6(1): 30-39. DOI: 10.1145/1007730.1007736.

[11] Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 2002; 16(1): 321-357.DOI: 10.1613/jair.953.

[12] Chawla NV, Lazarevic A, Hall LO, Bowyer KW. SMOTEBoost: Improving prediction of the minority class in boosting. In Book: Lavrač N, Gamberger D, Todorovski L, Blockeel H, eds. Knowledge discovery in databases. Berlin, Heidelberg, New York: Springer-Verlag; 2003: 107-119. DOI: 10.1007/978-3-540-39804-2_12.

[13] Fukunaga K. Introduction to Statistical pattern recognition. 2nd ed. San Diego: Academic Press; 1990.

[14] Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. Hoboken, NJ: Wiley-Interscience; 2000.

[15] Kryanev AV, Lukin GV. Mathematical methods for handling uncertain data [In Russian]. Moscow: "Fizmatlit" Publisher; 2003.

[16] Akimov AV, Donskikh AO, Sirota AA. Models and algorithms of digital image recognition under influence of warping and additive noise [In Russian]. Proceedings of Voronezh State University. Series: System Analysis and Information Technology 2018; 1: 104-118.

[17] Gramacki A. Nonparametric kernel density estimation and its computational aspects. Cham, Switzerland: Springer International Publishing AG; 2018: 42-49. ISBN: 978-3-319-71687-9.

[18] Dobrovidov AV, Ruds'ko IM. Bandwidth selection in nonparametric estimator of density derivative by smoothed cross-validation method. Automation and Remote Control 2010; 71(2): 209-224. DOI: 10.1134/S0005117910020050.
[19] Voronov IV, Mukhometzianov RN, Krasnova AA. Bandwidth selection in the approximation of probability density via Parzen-Rosenblatt method for small sample size [In Russian]. Radio Electronics Technology 2016; 1(9): 93-98.
[20] Donskikh AO, Minakov DA, Sirota AA. Optical methods of identifying the varieties of the components of grain mixtures based on using artificial neural networks for data analysis. Journal of Theoretical and Applied Information Technology 2018; 96(2): 534-542.

Authors' information

Alexander Anatolievich Sirota (1954) graduated from Voronezh State University in 1976 majoring in "Radiophysics and Electronics". Professor, Doctor of Technical Sciences (since 1995). Currently head of the Information Processing and Security Technologies chair at Voronezh State University. Research interests: analysis and design of information collection and processing systems, methods and techniques of information processes and systems computer modeling, system analysis in information security, digital image processing, neural networks and neural network technologies in decision-making systems. E-mail: *sir@cs.vsu.ru* .

Artem Olegovich Donskikh (1993) graduated from Voronezh State University in 2016 with master's degree in Information Systems and Technology. Currently postgraduate student at Information Processing and Security Technologies chair at Voronezh State University. Research interests: machine learning, digital image processing, data augmentation. E-mail: *a.donskikh@outlook.com* .

Alexey Viktorovich Akimov (1990) graduated from Voronezh State University in 2013 with master's degree in Information Systems and Technology. Currently postgraduate student at Information Processing and Security Technologies chair at Voronezh State University. Research interests: image recognition, machine learning.
E-mail: *akimov@rcnit.vsu.ru* .

Dmitry Anatolyevich Minakov (1982) graduated from Voronezh State University in 2005 majoring in "Optics". PhD in Physics and Maths (2008). Currently senior researcher at the physics laboratory of the Computer Science department of Voronezh State University. Research interests: machine learning, spectral analysis methods, fiber optic devices. E-mail: *minakov_d_a@mail.ru* .

Received March 15, 2019. The final version – April 10, 2019.