УДК 004.89
DOI 10.34822/1999-7604-2021-2-38-46

ПРАКТИЧЕСКИЙ ПОДХОД К ОПРЕДЕЛЕНИЮ ДОСТАТОЧНОСТИ ЭКСПЕРИМЕНТАЛЬНОЙ ВЫБОРКИ ВЕКТОРНЫХ СИГНАЛОВ ГАЗОАНАЛИТИЧЕСКОЙ МУЛЬТИСЕНСОРНОЙ ЛИНЕЙКИ ДЛЯ ОБУЧЕНИЯ ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ

В. С. Дыкин, В. Ю. Мусатов ✉, А. С. Варежников, В. В. Сысоев
Саратовский государственный технический университет им. Ю. А. Гагарина, Саратов, Россия
✉ *E-mail: vmusatov@mail.ru*

Предложен алгоритм нахождения достаточного объема репрезентативной совокупности данных выборки для обучения искусственных нейронных сетей по численным параметрам исходной выборочной совокупности на примере анализа векторного сигнала, генерируемого газоаналитической мультисенсорной линейкой на основе тонкой пленки SnO_2 при калибровке к воздействию CO, изопропанолу и этанолу в смеси с воздухом. В результате работы алгоритма найден минимальный необходимый объем обучающей выборки искусственных нейронных сетей, который позволяет достичь высокого (более 99 %) качества распознавания. Полученные результаты показывают работоспособность предложенного алгоритма.

Ключевые слова: распознавание образов, газовый сенсор, мультисенсорная линейка, анализ газов, обучающая выборка.

PRACTICAL APPROACH TO DETERMINING SUFFICIENCY OF EXPERIMENTAL SAMPLE OF GAS ANALYTICAL MULTISENSOR MICRO-ARRAYS VECTOR SIGNALS FOR TRAINING ARTIFICIAL NEURAL NETWORK

V. S. Dykin, V. Yu. Musatov ✉, A. S. Varezhnikov, V. V. Sysoev
Yuri Gagarin State Technical University of Saratov, Saratov, Russia
✉ *E-mail: vmusatov@mail.ru*

An algorithm for finding a sufficient amount of a representative set of sample data for training artificial neural networks using the numerical parameters of the original sample set is proposed. The algorithm is carried out on the example of analyzing a vector signal generated by a gas analytical multisensor micro-arrays based on a thin SnO_2 film when calibrated to the effect of CO, isopropanol, and ethanol in a mixture with air. As a result of the operation of the algorithm, the minimum required volume of the training sample of artificial neural networks was found, which allows achieving a high (more than 99 %) recognition quality. The results show the efficiency of the proposed algorithm.

Keywords: pattern recognition, gas sensor, multisensor micro-arrays, gas analysis, training set.

Введение

При решении задач распознавания образов с помощью методов искусственного интеллекта выбор данных для обучения является одним из определяющих факторов успешной классификации тестов [1]. В частности, когда используются искусственные нейронные сети (ИНС), обучающая выборка играет очень важную роль и должна отражать исследуемые свойства генеральной совокупности данных, т. е. быть репрезентативной [2]. Однако процесс сбора данных и создания выборки для обучения может быть слишком длительным, и получаемые данные могут быть несбалансированными [3]. Например, калибровка и сбор данных в газоаналитических системах, применяющих мультисенсорные линейки [4–5], зачастую требуют продолжительного времени с целью учета различных интерферирующих воздей-

ствий [6]. Поэтому требуются подходы к определению минимального объема выборки, обладающей свойством репрезентативности.

Можно отметить, что в общем случае репрезентативность выборочной совокупности определяется математически из информации о генеральной совокупности. Выборочная совокупность считается репрезентативной, если отклонение значения контролируемого признака от значения признака в генеральной совокупности не превышает в среднем 5 % [7]. Но в прикладных задачах, в том числе в газоанализаторах на основе мультисенсорных линеек, знания о генеральной совокупности априори недоступны [8]. Тем не менее, даже в условиях ограниченности исходных данных, обучение ИНС должно быть выполнено корректно, чтобы выходной сигнал сети в условиях тестирования незначительно отличался от калибровочных примеров. Другими словами, обучающая выборка должна быть репрезентативной, чтобы представлять обобщающую способность для ИНС. Считается, что для хорошего обобщения достаточно, чтобы размер обучающего множества N удовлетворял следующему соотношению [9]:

$$N = O(W / \varepsilon), \tag{1}$$

где W – общее количество свободных параметров (синаптических весов и порогов) ИНС;

ε – допустимая точность ошибки классификации;

$O(\ )$ – порядок величины, заключенной в скобки. Например, для ошибки в 10 % количество примеров должно в 10 раз превосходить количество свободных параметров сети.

Величина, определяемая выражением (1), позволяет оценить размер обучающей выборки приблизительно, на уровне определения его порядка, что недостаточно для практического применения, поэтому для определения размера обучающего множества используют другие подходы. Один из таких подходов основывается на рассмотрении способности ИНС к обобщению. Такая способность определяется тремя факторами: размером обучающего множества и его репрезентативностью, архитектурой ИНС и физической сложностью рассматриваемой задачи. Причем, как показано в [10], критичным при обучении является не размер множества вычисляемых ИНС функций отображения входа на выход, а VC-измерение сети, определяющее емкость алгоритма классификации и показывающее максимальное число бинарных образов, которые алгоритм может корректно разделить [11]. В этом случае считают, что существует такая константа K, при которой достаточным размером обучающего множества для любого алгоритма является:

$$N = \frac{K}{\varepsilon}\left(\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right), \tag{2}$$

где ε – допустимая точность ошибки распознавания; δ – доверительный интервал.

Выражение (2) часто применяют к обучению ИНС с учителем независимо от типа алгоритма обучения. Однако значения достаточного обучающего множества, полученные на основе вычисления VC-измерения, могут существенно расходиться с экспериментальными данными [12] вследствие независимости от распределения и пессимистического характера теоретических оценок. Более того, VC-измерение является комбинаторным понятием и не связано с геометрическим понятием измерения. Другими словами, количество примеров, необходимых для обучения системы данным некоторого класса, строго пропорционально VC-измерению этого класса. Это означает, что VC-измерение относится только к обучаемой системе и неприменимо для определения параметров выборочной совокупности.

Другой подход к достаточности выборки предложен в социологической литературе, где при опросах применяют следующий формальный критерий: отбор респондентов производят из целевой группы согласно социально-демографическим характеристикам генеральной совокупности на основе выбора экспертов, социологов-практиков [2]. Например, когда генеральная совокупность превышает 5 000 человек, эмпирическая выборка составляет не более 10 % от генеральной совокупности [11, 13].

Таким образом, анализ литературы показывает, что не существует строгих и однозначных критериев, а также алгоритмов поиска минимального объѐма выборки данных, обеспечивающей качественное обучение ИНС. Более того, часто отмечается, что критериями достаточного минимального объема выборочного множества являются числовые усредненные параметры контрольных признаков элементов генеральной совокупности и их дисперсии: как правило, чем больше дисперсия, тем больше должен быть объем выборочной совокупности [7–8]. Рассмотрим алгоритм нахождения достаточного объема репрезентативной совокупности по численным параметрам исходной выборочной совокупности на примере анализа векторного сигнала, генерируемого газоаналитической мультисенсорной линейкой, при калибровке к воздействию тестовых газов.

Материал и методы

Газоаналитическая мультисенсорная линейка и векторный сигнал

В работе рассмотрена газоаналитическая мультисенсорная линейка в виде чипа с матрицей нановолокон оксида олова, генерирующих хеморезистивный сигнал к различным газам при нагреве до 350 °С [14]. Внешний вид чипа и архитектура мульти-электродов представлены на рис. 1 (а–г). Оксидные нановолокна были нанесены на кремниевые подложки (с изолирующим оксидным слоем на поверхностях) площадью 10 × 10 мм^2 [15]. Подложка содержит компланарные электроды на фронтальной стороне и нагревательные элементы на обратной. Архитектура электродов позволяет иметь на чипе набор хеморезисторов, совокупный векторный сигнал которых служит для селективного анализа тестовых газов после обработки методами распознавания образов [16].

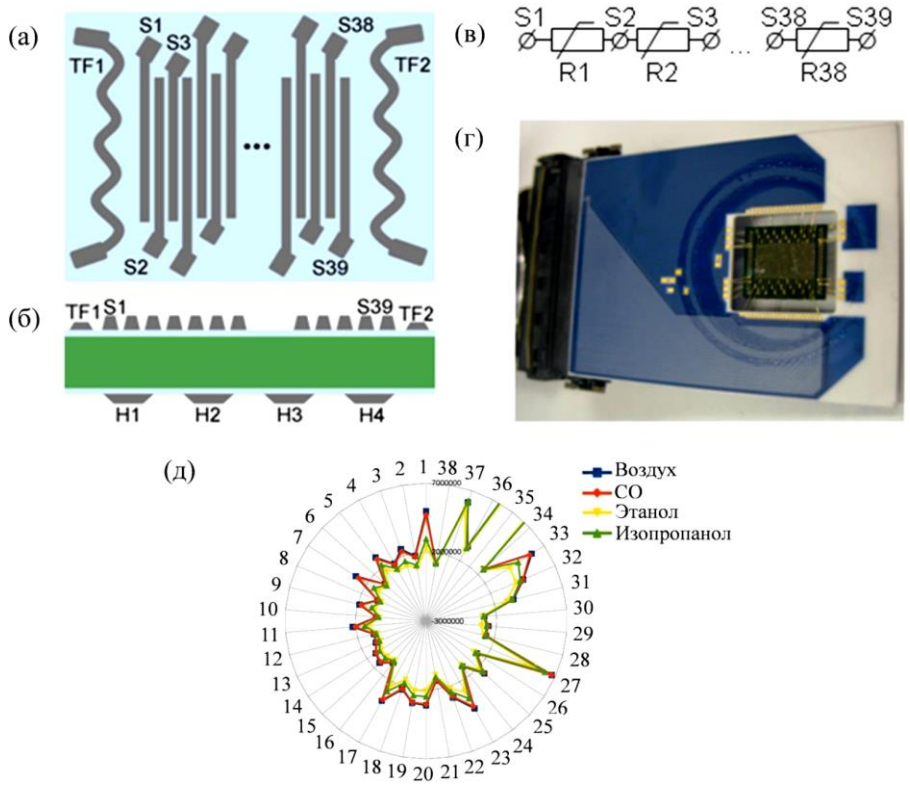

Рис. 1. Однокристальный мультисенсорный чип с матрицей нановолокон оксида олова:
а) фронтальный вид мультисенсорного чипа (TF1–2 – терморезисторы для контроля температуры, S1–S39 – платиновые компланарные электроды); *б)* поперечный разрез чипа (H1–H4 – нагреватели); *в)* электрическая схема мультисенсорного чипа; *д)* пример векторного сигнала газоаналитической мультисенсорной линейки, реализованной на чипе, в атмосфере чистого воздуха, при воздействии примесей CO и паров спиртов (этанола, изопропанола)
Примечание: составлено авторами.

В данной работе был исследован векторный сигнал такого мультисенсорного чипа, полученный при воздействии CO, этанола и изопропанола в смеси с воздухом. Пример векторного сигнала, генерируемого данной мультисенсорной системой, состоящей из 38 сенсоров, представлен на рис. 1д. Для обработки векторного сигнала чипа использовалась ИНС, для обучения которой применяли выборку данных – векторных сигналов, состоящую из 586 измерений, соответствующих 3 выходным классам.

Искусственная нейронная сеть

В данной работе применялась ИНС прямого распространения с обратным распространением ошибки, состоящая из одного скрытого слоя с количеством нейронов, равным 10 (рис. 2). Для обучения применялся метод Левенберга – Марквардта.

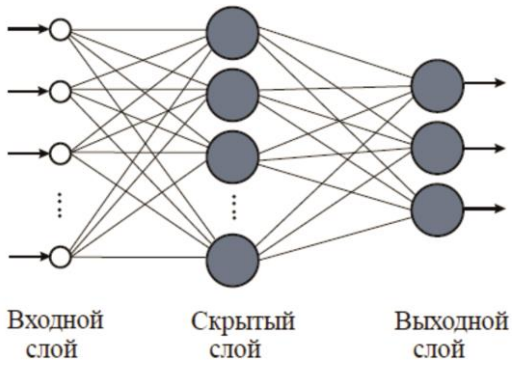

Рис. 2. Схема ИНС, использованная в работе
Примечание: составлено авторами.

Число нейронов во входном слое ИНС соответствовало числу компонент мультисенсорного векторного сигнала, а количество выходных нейронов – количеству тестовых газовых смесей (или распознаваемых классов), равному 3. Сигналы нейронов выходного слоя были заданы так, что каждому классу соответствовал один нейрон; сигнал нейрона, равный 1, означал наличие соответствующего тестового газа, 0 – его отсутствие.

Результаты

Было предположено наличие зависимости репрезентативности выборочной совокупности данных от значений и скорости изменения статистических параметров совокупности – математического ожидания и дисперсии. С целью нахождения этой зависимости варьировали значения статистических параметров исходной совокупности и анализировали изменения результативности распознавания ИНС в предположении, что результативность распознавания при малых объемах обучающей совокупности линейно зависит от репрезентативности этой совокупности. Для расширения обучающей выборки был выбран алгоритм моделирования коррелированных данных следующего вида [10]:

1) определение нормально распределенного случайного значения из эмпирической выборки (независимой величины x);

2) определение математического ожидания $m_{y/x}$ зависимой величины y:

$$m_{y/x} = m_y + q \cdot \sigma_y / \sigma_x \cdot (x - m_x),$$

где q – коэффициент корреляции между x и y;

3) определение среднеквадратичного отклонения $\sigma_{y/x}$ зависимой величины y:

$$\sigma_{y/x} = \sigma_y \cdot \sqrt{(1 - q^2)};$$

4) определение зависимого случайного числа y в предположении о его нормальном распределении:

$$y = randn \cdot \sigma_{y/x} + m_{y/x},$$

где *randn* – случайное нормально-распределенное число.

Качество распознавания рассчитывалось как отношение правильно распознанных измерений к их общему числу.

На рис. 3 показаны результаты анализа зависимости качества распознавания ИНС классов, соответствующих тестовым аналитам (этанол, изопропанол, CO), от отношения математических ожиданий M/M_0, при изменении математического ожидания и дисперсии обучающей выборки. Изменение математических ожиданий происходило пропорционально разности среднего математического ожидания по всем классам и среднего математического ожидания каждого класса для каждого сенсорного элемента в мультисенсорной линейке:

$$M_j = M_j + k \cdot (\bar{M} - M_j). \quad (3)$$

При этом отношение $M/M_0 = 1$ означает соответствие математическим ожиданиям исходной выборки без изменений, а $M/M_0 = 0$ соответствует случаю, когда математические ожидания каждого класса совпадают со средним математическим ожиданием всех классов по каждому сенсорному элементу мультисенсорной линейки. Кривые соответствуют различным изменениям дисперсии D от исходного значения в выборке D_0 в диапазоне $D/D_0 = [0{,}5; 2]$.

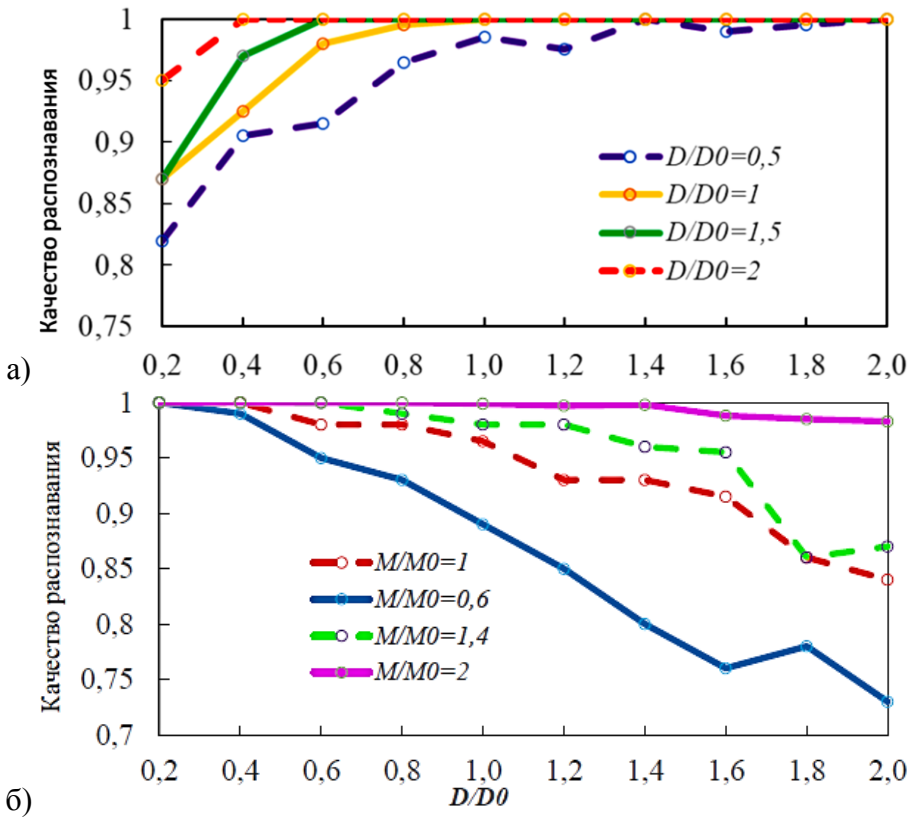

Рис. 3. Зависимость результата распознавания тестовых паров с помощью ИНС при анализе выборки векторных сигналов мультисенсорной линейки во время ее модификации изменением математического ожидания (а) и изменением дисперсии данных в выборке (б). Кривые соответствуют различным изменениям дисперсии D от исходного значения в выборке D_0 в диапазоне $D/D_0 = [0{,}5;2]$

Примечание: составлено авторами.

Кривые на рис. 3а показывают существенную зависимость качества распознавания от изменения математических ожиданий. Эта зависимость увеличивается с увеличением дисперсии относительно D_0. Так, при $D/D_0 = 0{,}5$ качество распознавания с изменением математического ожидания M изменяется на 5 % (от 0,95 до 1), а при $D/D_0 = 2$ качество распознавания изменяется в более широком диапазоне: от 0,82 до 1. Исходя из геометрической интерпретации данных, видно, что более значимые изменения математических ожиданий ведут к более существенному отличию тестовых классов в выборке исходных данных. Причем уменьшение дисперсии данных ведет к более качественному распознаванию, которое быстрее достигается с изменением математического ожидания. Поэтому можно сделать вывод о прямой зависимости результативности классификации классов с помощью ИНС от различия классов данных в выборке.

На рисунке 3б представлена зависимость качества распознавания ИНС от изменения дисперсии каждого класса обучающих данных при различных значениях изменения математических ожиданий. Как видно из этого рисунка, наблюдается обратная зависимость качества распознавания от изменения дисперсии: чем выше дисперсия, тем хуже распознавание, что в некоторой степени является ожидаемым результатом. Причем эта зависимость тем сильнее, чем меньше значение изменения M/M_0. При большем значении M/M_0, например равном 2, изменение качества распознавания составляет меньше 2 % (от 100 % до 98,3 %). Можно отметить, что изменение дисперсии существенно влияет на результативность распознавания ИНС при малых различиях данных в выборке, относящихся к разным классам, и практически не влияет, когда центры классов находятся на больших расстояниях друг от друга.

Для анализа зависимости качества распознавания от соотношения межклассовой D_{class} и внутриклассовой D_{inter} дисперсии данных в выборке был применен дисперсионный анализ [10–11], в котором межклассовая дисперсия определяется расположением классов, а внутриклассовая дисперсия определяется случайными шумами и неучтенными факторами. Были составлены 30 обучающих выборок размерностью в 45 векторных сигналов мультисенсорной линейки, по 15 сигналов, соответствующих каждому классу тестового аналита, и проведена оценка их распознавания с помощью разработанной ИНС. На рис. 4 показана зависимость качества распознавания от соотношения межклассовой и внутриклассовой дисперсии D_{class}/D_{inter}. Из полученных результатов следует, что при изменении D_{class}/D_{inter} от 2,6 до 4,6 распознавание слабо зависит от соотношения межгрупповой и внутригрупповой дисперсии данных в выборке. Это, по-видимому, связано с достаточно большим отношением указанных дисперсий в этом диапазоне и существенным превышением вариаций, вызванных аналитами, над «внутриклассовым» шумом.

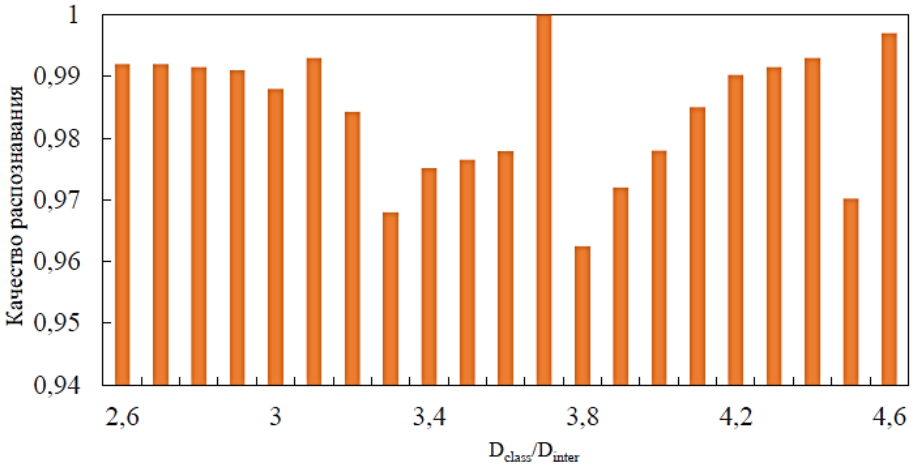

Рис. 4. Зависимость качества распознавания от соотношения межгрупповой и внутригрупповой дисперсии, D_{class}/D_{inter}

Примечание: составлено авторами.

Учитывая недостаток дисперсионного анализа применительно к решаемой задаче, необходимо ввести формализацию границ классов данных. С этой целью была предложена мера C, которая заключается в минимальном для всех классов отношении разности математических ожиданий к полусумме дисперсий классов:

$$C = \{| M_i - M_j | / 0,5(D_i + D_j)\}, \quad (4)$$

где i, j – номера классов;

M_i, M_j, D_i, D_j – соответствующие математические ожидания и дисперсии (рис. 5). Значение меры C рассчитывается так, что если $C > 1$, то классы n_j не пересекаются, а если $C \leq 1$, то классы пересекаются. Причем значение C рассчитывается для каждого сенсора мультисенсорной линейки. Это позволяет учесть нечувствительность распознавания ИНС к пересечению двух классов при достаточно удаленном третьем классе.

Следует отметить, что в этом случае анализ большого количества величин C для всей мультисенсорной линейки представляет довольно трудоемкую задачу. Поэтому для упрощения был выявлен критерий сенсорного сигнала, который в наибольшей степени влияет на результат распознавания ИНС. С этой целью статистические параметры сенсорных сигналов в выборке (дисперсия и математическое ожидание) изменялись последовательно, в зависимости от величины C, и наблюдалось их влияние на качество распознавания.

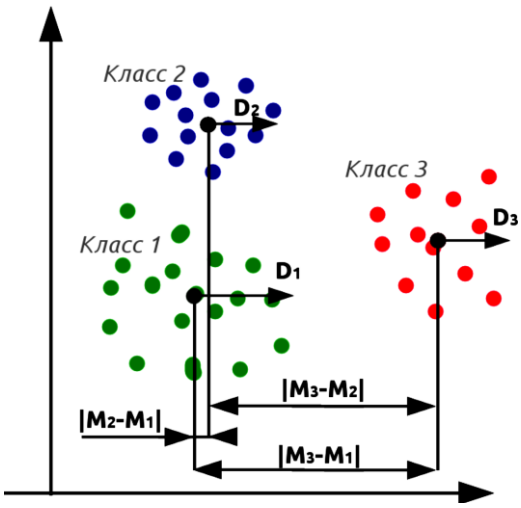

Рис. 5. Схематичное представление классов в *N*-мерном пространстве исходных векторных мультисенсорных сигналов для введения их формальных границ, где *N* – количество компонент вектора или количество сенсоров в мультисенсорной линейке. Для иллюстрации показаны 2-мерные векторные сигналы. M_i, D_i – математическое ожидание и дисперсия в данных выборки, соответствующих каждому классу *i*

Примечание: составлено авторами.

В результате был составлен следующий алгоритм поиска достаточности выборки для обучения ИНС:

1) определение сенсора в мультисенсорной линейке, у которого наблюдается наибольшее значение величины C;

2) определение допустимого интервала изменения дисперсии каждого класса данных в обучающей выборке:

$$D_{\min} = D - (0,5cD); \quad D_{\max} = D + (0,5cD).$$

Предполагается, что дисперсия данных в обучающей выборке при проведении n-итераций лежит внутри допустимого интервала. В случае, если величина дисперсии выхо-

дит за указанные пределы либо наибольшее значение величины *C* появилось у другого сенсора в мультисенсорной линейке, алгоритм необходимо выполнить еще раз. Одна итерация алгоритма соответствует увеличению выборки на один пример каждого класса.

Моделирование работы алгоритма, полученного на основе анализа вышеописанных данных, было проведено в оболочке MatLab®, вер. 8.1 (MathWorks, Inc, США), при обработке мультисенсорного сигнала, полученного к более сложным аналитам – ароматам, генерируемым различными алкогольными напитками (бренди (40 об. % спирта), вино (12 об. % спирта), шампанское (12 об. % спирта)) с целью их селективного анализа. Для того, чтобы изменение содержания этилового спирта не вносило вклад в распознавание, все образцы были приведены к 11 об. % спирта путем соответствующего разбавления. Размер обучающей выборки полученных мультисенсорных векторных сигналов чипа составлял 75 (3 класса по 25 экспозиций). На рис. 6 представлены результаты распознавания этих данных после обработки ИНС при последовательном увеличении размера обучающей выборки. Как видно из рисунка, при размере выборки, превышающей 15 экспозиций, распознавание превышает 99 %, что вполне достаточно на практике.

Рис. 6. Результаты распознавания векторных сигналов газоаналитической мультисенсорной линейки с помощью ИНС при последовательном увеличении обучающей выборки по предложенному алгоритму
Примечание: составлено авторами.

Более того, значение параметра *n*, равное 5, оказывается достаточным для построения выборки, подходящей для обучения ИНС с качеством распознавания, близким к 100 %. Например, минимальный необходимый объем обучающей выборки ИНС равен 22 экспозициям каждого тестового аромата, при котором качество распознавания ИНС составляет 99,98 %.

Обсуждение и заключение

Таким образом, предложен алгоритм нахождения достаточного объема репрезентативной совокупности данных выборки для обучения ИНС по численным параметрам исходной выборочной совокупности на примере анализа векторного сигнала, генерируемого газоаналитической мультисенсорной линейкой при калибровке к воздействию тестовых газов.

Работа алгоритма экспериментально проверена для данных, полученных с газоаналитического мультисенсорного чипа на основе тонкой пленки SnO_2 при экспонировании к ряду практически важных аналитов. В результате работы алгоритма был найден минимальный необходимый объем обучающей выборки ИНС, который позволяет достичь высокого (более 99 %) качества распознавания. Полученные результаты показывают работоспособность предложенного алгоритма.

Благодарность

Авторы благодарят за сотрудничество при изготовлении мультисенсорного чипа д-ра М. Зоммера (Технологический Институт Карлсруэ, Германия). Исследование выполнено при поддержке Минобрнауки РФ в рамках госзадания СГТУ им. Ю. А. Гагарина.

Литература

1. Kaltenecker C., Grebhahn A., Siegmund N., Apel S. The Interplay of Sampling and Machine Learning for Software Performance Prediction // IEEE Software. 2020. DOI 10.1109/MS.2020.2987024.

2. Ильясов Ф. Н. Репрезентативность результатов опроса в маркетинговом исследовании // Социологические исследования. 2011. № 3. С. 112–116.

3. Fernández A., García S., Herrera F. Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution // Lecture Notes in Computer Science. 2011. P. 1–10. DOI 10.1007/978-3-642-21219-2_1.

4. Мусатов В. Ю., Сысоев В. В. Обработка данных мультисенсорных систем «электронный нос» на основе методов искусственного интеллекта // Системы искусственного интеллекта в мехатронике / под ред. Большакова А. А., Бровковой М. Б., Глазкова В. П. и др. Саратов : Сарат. гос. тех. ун-т. 2015. С. 146–184.

5. Hierlemann A., Gutierrez-Osuna R. Higher-Order Chemical Sensing // Chem Rev. 2008. No. 108. P. 563–613.

6. Kiselev I., Sysoev V., Kaikov I., Koronczi I., Tegin R. A. A., Smanalieva J., Sommer M., Ilicali C., Hauptmannl M. On Temporal Stability of Analyte Recognition with E-nose Based on Metal Oxide Sensor Array in Practical Applications // Sensors. 2018. Vol. 18, P. 22.

7. Зайцев Г. Н. Математическая статистика в экспериментальной ботанике. М. : Наука. 1984. 425 с.

8. Хайкин С. Нейронные сети / пер. с англ. М. : Вильямс, 2006. 1104 с.

9. Widrow B., Steams S. D. Adaptive Signal Processing. Englewood Cliffs, NJ : Prentice Hall, 1985. 475 p.

10. Шелухин О. Моделирование информационных систем. М. : Горячая линия – Телеком, 2011. X с.

11. Гайдышев И. Анализ и обработка данных : спец. справочник. СПб. : Питер, 2001. 750 с.

12. Vapnik V., Levin E., Cun Y. L. Measuring the VC-dimension of a Learning Machine // Neural Computation. 1994. Vol. 6. P. 851 –876.

13. Горшков М. К., Шереги Ф. Э. Прикладная социология: методология и методы. М. : Альфа ; ИНФРА-М, 2009. 416 с.

14. Sysoev V. V., Goschnick J., Schneider T., Strelcov E., Kolmakov A. A Gradient Microarray Electronic Nose Based on Percolating SnO_2 Nanowire Sensing Elements // Nano Letters. 2007. Vol. 7, Iss. 10. P. 3182–3188. DOI 10.1021/nl071815+.

15. Sysoev V. V., Strelcov E., Kar S., Kolmakov A. The Electrical Characterization of a Multi-electrode Odor Detection Sensor Array Based on the Single SnO_2 Nanowire // Thin Solid Films. 2011. Vol. 520. P. 898–903. DOI 10.1016/j.tsf.2011.04.179.

16. Sysoev V. V., Strelcov E., Kolmakov A. Multisensor Micro-Arrays Based on Metal Oxide Nanowires for Electronic Nose Applications. Metal Oxide Nanomaterials for Chemical Sensors. Springer : New York, Heidelberg, Dordrecht, London, 2013. Chapter 15. P. 465–502.