Математические методы исследования
Mathematical methods of investigation

DOI: https://doi.org/10.26896/1028-6861-2023-89-7-71-77

ПРОЦЕДУРА ПРОВЕРКИ ОДНОРОДНОСТИ ВЫБОРОК ТЕКСТОВЫХ ДОКУМЕНТОВ НА ОСНОВЕ НЕПАРАМЕТРИЧЕСКИХ КРИТЕРИЕВ

© **Шахим Ильмирович Сафин, Владимир Олегович Толчеев***

Национальный исследовательский университет «Московский энергетический институт», Россия, 111250, Москва, ул. Красноказарменная, д. 14; *e-mail: tolcheevvo@mail.ru

Статья поступила 10 января 2023 г. Поступила после доработки 27 февраля 2023 г. Принята к публикации 30 марта 2023 г.

При построении высокоточных классификаторов одной из важнейших задач является формирование достаточно больших репрезентативных и непротиворечивых выборок. В частности, при анализе и обработке текстовых документов объединяют наборы данных, полученных из различных информационных источников. В ряде случаев из-за нехватки профильных текстов на русском языке датасет расширяют за счет добавления переведенных англоязычных документов. В таких ситуациях целесообразно оценивать однородность-неоднородность объединяемых массивов. Однако подобная проверка осложняется тем, что документы представляют собой многомерные векторы, корректное сопоставление которых является весьма нетривиальной задачей. Недостаточная разработанность процедур проверки однородности выборок для многомерного случая приводит к тому, что на практике проблема возможных различий в данных игнорируется как несущественная. Как следствие, обучение классификаторов проводится по выборкам, представляющим собой смесь достаточно разнотипных текстов, и результирующее качество категоризации не улучшается (или даже ухудшается). Все это обуславливает актуальность разработки процедуры проверки однородности документальных выборок. Для этого авторы провели комплексное изучение проблемы сдвига в текстовых данных, выявили и проанализировали причины, которые определяют неоднородность документальных массивов. Исследуемые выборки состоят из библиографических описаний научных статьей (название, аннотация, ключевые слова). Авторы разработали процедуру оценки однородности двух выборок, имеющих приблизительно одинаковый объем и единый способ расчета весов терминов. Для сопоставления использовали центроиды, которые имеют размер общего словаря двух датасетов (в случае отсутствия некоторых терминов в соответствующие позиции центроидов проставляют нулевые значения). Представление выборок в виде «терминологических портретов» (центроидов) позволяет свести проверку однородности многомерных векторов-документов к хорошо изученной задаче анализа двух одномерных связанных выборок, для решения которой применяли непараметрические критерии (в частности, критерий знаков и критерий знаковых рангов Вилкоксона). Предложенная процедура проверки однородности выборок на основе непараметрических критериев проверена на трех коллекциях документов, полученных из русско- и англоязычных источников.

Ключевые слова: интеллектуальный анализ текстовых данных; однородность-неоднородность выборок; непараметрические критерии; сравнение центроидов.

PROCEDURE FOR CHECKING THE UNIFORMITY OF SAMPLES OF TEXT DOCUMENTS BASED ON NONPARAMETRIC CRITERIA

© **Shahim I. Safin, Vladimir O. Tolcheev***

National Research University “Moscow Power Engineering Institute”, 14, Krasnokazarmennaya ul., Moscow, 111250, Russia; *e-mail: tolcheevvo@mail.ru

Received Junuary 10, 2023. Revised February 27, 2023. Accepted March 30, 2023.

One of the most important tasks in Text Mining is the formation of sufficiently large representative and consistent samples (datasets). Usually, datasets are obtained from various information sources. In some

cases, due to the lack of specialized texts in Russian, the dataset is expanded by adding translated English-language documents. In such situations, it is advisable to evaluate the uniformity-heterogeneity of the combined arrays. However, such a verification is complicated by the fact that the documents are multidimensional vectors, the correct comparison of which is a very non-trivial task. Insufficient elaboration of procedures for checking the uniformity of samples for the multidimensional case leads to the fact the problem of possible differences in data is ignored that in practice as insignificant. As a result, classifiers are trained on samples that are a mixture of quite diverse texts, and the resulting quality of categorization does not improve (or even deteriorates). Thus, it seems relevant to develop a procedure for checking the uniformity of documentary samples. To do this, we provide a comprehensive study of the problem of shift in textual data, identified and analyzed the reasons that cause the heterogeneity of documentary arrays. In this study, the datasets consist of bibliographic descriptions of scientific articles (title, abstract, keywords). The authors develop a procedure for assessing the homogeneity of two samples having approximately the same volume and the same method for calculating the weights of terms. For comparison, centroids are used, which have the size of a common dictionary of two datasets (in the absence of some terms, zero values are put in the corresponding positions of the centroids). The representation of samples in the form of "terminological portraits" (centroids) allowed us to reduce the verification of the homogeneity of multidimensional document vectors to a well-studied problem of analyzing two one-dimensional connected samples, for which nonparametric criteria were used. The sign criterion and the Wilcoxon sign rank criterion were used in the study. The proposed procedure for checking the uniformity of samples was tested on three collections of documents obtained from Russian and English-language sources.

Keywords: Text Mining; uniformity-heterogeneity of samples; nonparametric criteria; comparison of centroids.

Введение

В задачах прикладной статистики и машинного обучения одним из ключевых моментов построения высокоточных классификаторов является формирование достаточно больших репрезентативных и непротиворечивых выборок, содержащих актуальную информацию, — датасетов, набора (массив, коллекция) данных. При этом чем больше обучающих примеров имеется у исследователя, тем сложнее модели можно построить, достигнув лучших показателей качества [1, 2].

Однако увеличение объема обучающей выборки не всегда приводит к улучшению результатов классификации. В ряде случаев они практически не меняются или даже несколько ухудшаются. В частности, такой эффект наблюдается при анализе документальных датасетов по научным тематикам. Для увеличения объема обучающей выборки и более полного охвата предметной области обычно проводят объединение (независимых) подмножеств документов, полученных из различных информационных ресурсов (цифровых библиотек, сайтов специализированных журналов, конференций, профессиональных ассоциаций) [3]. При расширении выборки за счет данных из других источников редко оценивают степень однородности агрегируемых текстовых массивов. Как следствие, обучение зачастую проводят по неоднородным выборкам, представляющим смесь разнотипных публикаций. Такая ситуация во многом сложилась из-за отсутствия общепринятой и просто реализуемой процедуры, позволяющей определять различия при объединении нескольких датасетов [4].

Конкретизируем понятие однородности текстовых выборок. Это означает, что все документы, извлеченные из разных хранилищ данных, принадлежат одной генеральной совокупности с общим законом распределения вероятностей [1]. Пусть имеются две выборки, полученные из разных источников, и для их возможной консолидации проводится проверка на однородность.

Предположим, что первая выборка X_1, ..., X_j, ..., X_{N1} имеет функцию распределения $F(X)$, а вторая Y_1, ..., Y_k, ..., Y_{N2} — функцию распределения $G(X)$. Сформулируем нулевую и альтернативные гипотезы [1]:

$$H_0: F(X) = G(X) \text{ при всех } X,$$

$$H_1: F(X_0) \neq G(X_0) \text{ хотя бы при одном } X = X_0.$$

На практике для оценки однородности обычно используют характеристики положения и разброса (матожидание, медиану, дисперсию) и непараметрические критерии, не требующие знания закона распределения. В рамках прикладной статистики для этих целей разработаны специальные критерии «сдвига» (определение различий в характеристиках положения) и «масштаба» (определение различий в степени рассеивания случайных величин) [1, 5]. Именно данный подход, как представляется, является наиболее подходящим способом построения процедуры проверки однородности выборок, полученных с разных информационных ресурсов.

Анализ причин неоднородности текстовых документов

При решении реальных задач важно не только определить степень однородности выборок, но и понять причины возникновения неоднородности и возможные способы ее устранения. Это

имеет особое значение при построении классификаторов, результатом работы которых должно стать получение как можно более однородных подмножеств документов (классов) [1, 6].

Проанализируем более подробно возможные причины появления неоднородности в текстовых коллекциях. Пусть имеются две одноклассовые выборки — X_1, ..., X_j, ..., X_{N1} и Y_1, ..., Y_k, ..., Y_{N2}, полученные из различных источников по одной тематике (ключевым словам) и имеющие приблизительно одинаковый объем $N_1 \approx N_2$. При этом каждый документ описывается многомерным вектором.

Рассмотрим вероятность совместного появления двух случайных величин — X или Y (документ) и Q (класс): $P(Q, X) = P(Q)P(X/Q)$ и $P(Q, Y) = P(Q)P(Y/Q)$.

В современной литературе выделяют различные аспекты проблемы однородности-неоднородности в зависимости от исследуемых предметных областей (медицинская диагностика, розничная торговля, продажа недвижимости, выдача кредитов) [7 – 12]. Анализ документальных датасетов имеет свою специфику [13 – 15]. На наш взгляд, целесообразно рассматривать следующие основные причины неоднородности, возникающие в интеллектуальном анализе текстовых данных (Text Mining).

I. Терминологические различия («сдвиг в терминах»). При наличии такого сдвига одинаковые термины ($x^{(i)}$ и $y^{(i)}$) в двух выборках по-разному проявляют себя. В одной выборке они играют «лидирующую» роль (большое значение $P(x^{(i)})$), в другой — незначительную ($P(y^{(i)})$ мало).

Такая ситуация может возникнуть, если информационные источники, из которых извлекаются тексты, узко специализированные и содержат документы только по некоторым тематическим направлениям предметной области. Например, если класс Q представляет собой научные труды по интеллектуальному анализу данных, то в $\{X_j\}$ могут доминировать публикации по обучению ансамблей классификаторов, а в $\{Y_j\}$ — статьи по использованию нейросетевых моделей.

II. Неоднозначная трактовка тематики Q в разных хранилищах данных («сдвиг тематики», «сдвиг метки»). Такой сдвиг часто случается при анализе текстовых документов, которые сложно отнести к одной метке (это также характерно для быстро развивающихся и междисциплинарных областей). Несмотря на равенство условных вероятностей ($P(Q/X) = P(Q/Y)$) эксперты относят документ к неэквивалентным классам (например, один и тот же документ в разных цифровых библиотеках может быть отнесен к рубрикам «Машинное обучение», «Информационный поиск», «Рекомендательные системы»).

III. «Дрейф концепции» — одновременное изменение $P(Q/X)$ и $P(Q/Y)$. Он указывает на динамические процессы в предметной области. В случае анализа научных тематик наличие дрейфа чаще всего свидетельствует о переосмыслении прежней концепции или, возможно, формировании нового направления (класса). Например, в течение последнего десятилетия из рубрики «Машинное обучение» выделилось самостоятельное направление «Глубокое обучение».

Эти причины в той или иной степени присутствуют при решении большинства практических задач в области Text Mining и совсем не обязательно приводят к сдвигу в данных, оказывая негативное влияние на качество обучения и классификации (чаще всего небольшой сдвиг не приводит к искажению решающего правила и границ между классами) [16].

Необходимо также отметить, что многие различия в данных сглаживаются на больших выборках. Если исследователь имеет в своем распоряжении огромные массивы документов (Big Data) и значительные вычислительные ресурсы, то неоднородность текстов нивелируется благодаря широкому охвату источников.

Однако достаточно часто исследователь не имеет доступа к Big Data и высокопроизводительному оборудованию. В этом случае один из способов увеличения качества классификации заключается в формировании большой выборки из разных информационных ресурсов с контролем однородности добавляемых подмножеств документов. Возникает необходимость разработки процедуры обнаружения различий в текстовых коллекциях, которая позволит судить об их однородности-неоднородности.

Для устранения каждой из вышеуказанных причин неоднородности текстовых данных требуется применение отдельной (специализированной) технологии. Так, обнаружение «сдвига тематики» может быть достигнуто путем укрупнения рубрик или введения более сложной «мягкой» классификации с простановкой нескольких меток для каждого документа. Если при эксплуатации классификатора появляется «дрейф концепции», то наиболее эффективным средством является дообучение (переобучение) классификатора на актуальных документальных массивах. Наименее формализованы, на наш взгляд, анализ терминологической близости разных датасетов и оценка их однородности.

Процедура проверки однородности выборок текстовых документов с помощью непараметрических критериев

В данной работе предлагаем использовать центроид («терминологический портрет») в ка-

честве типичного представителя каждого класса. Для корректного сравнения центроидов необходимо, чтобы исследуемые одноклассовые выборки имели (почти) одинаковый объем.

Проверка однородности двух наборов текстов сводится к анализу так называемых связанных выборок [1, 17], где вместо самих датасетов используются их центроиды. Эти центроиды имеют размер общего словаря двух датасетов. В случае отсутствия в одном из датасетов некоторых терминов веса соответствующего центроида принимают нулевые значения.

Построение центроидов также позволяет провести предварительный анализ числа несовпадающих терминов. Для этого можно рассчитать, например, коэффициент Жаккара

$$J = \frac{A}{A + B + C}, \quad (1)$$

где A — число терминов, общих для двух выборок; B — число терминов, встречающихся только в первой выборке; C — число терминов, встречающихся только во второй выборке.

Достаточно малые значения коэффициента Жаккара свидетельствуют о существенных терминологических различиях между центроидами, при которых нецелесообразно проведение дальнейших исследований на однородность. В случае достаточно высоких значений коэффициента Жаккара имеет смысл более детально проанализировать однородность выборок с помощью непараметрических статистических критериев [1, 18 – 21].

Рассмотрим процедуру применения непараметрических критериев для проверки однородности (идентичности) терминологического состава наборов данных. Имеются две одноклассовые выборки текстовых документов приблизительно одинакового размера ($N_1 \approx N_2$): X_1, ..., X_j, ..., X_{N1} и Y_1, ..., Y_k, ..., Y_{N2}. В случае двух многоклассовых выборок, которые обычно исследуются в Text Mining, необходимо отдельно оценивать однородность каждого класса. Например, если в двух сравниваемых выборках имеется класс «Интеллектуальный анализ данных» (ИАД) и мы хотим проверить степень однородности документов по этой тематике, то требуется изучить класс ИАД отдельно от других классов. В этом случае выборка $\{X_j\}$ будет соответствовать текстам по тематике ИАД в первой выборке, а $\{Y_j\}$ — этой же тематике во второй выборке.

Допустим, что словарь первой выборки имеет размер M_1, а словарь второй выборки — M_2. Проверка однородности выборок проводится в общем признаковом пространстве размера M ($M = M_1 + M_2$), т.е. M — число оригинальных терминов, которые более одного раза встретились в документах $\{X_j\}$ и $\{Y_j\}$ (предполагается, что ранее были удалены стоп-слова и проведена лемматизация). Таким образом, центроиды имеют размер M, в случае отсутствия термина в одной из выборок в соответствующую позицию центроида записывается нулевое значение. Заполнение нулями позиций отсутствующих терминов не нарушает важного предположения о непрерывности функции результатов наблюдения, на которое требуется обращать внимание при использовании непараметрических (ранговых) критериев, так как их введение не приводит к появлению совпадающих значений и «слипанию» наблюдений в связанных выборках [1, 17].

Элементы центроидов вычисляют по формулам:

$$C_1^{(i)} = \frac{1}{N_1}\sum_{j=1}^{N_1} X_j^{(i)},\ C_2^{(i)} = \frac{1}{N_2}\sum_{j=1}^{N_2} Y_j^{(i)}.$$

где $X_j^{(i)}$ и $Y_j^{(i)}$ — частоты появления i-го термина в j-х документах первой и второй одноклассовых выборок.

Термины в первом центроиде упорядочиваются от самых больших весов к самым малым (нулевым) — $C_1^{(1)}$, ..., $C_1^{(i)}$, ..., $C_1^{(m)}$, термины во втором центроиде повторяют порядок следования терминов первого центроида. Если $C_1^{(i)}$ — вес слова «интеллектуальный» в первой выборке, то $C_2^{(i)}$ — вес этого же слова во второй выборке. Таким образом, как и требуется при анализе связанных выборок, над каждым признаком (термином) проводят два «измерения». Одно показывает средний вес этого термина в первой выборке, а другое — средний вес во второй выборке.

Данный подход позволяет анализировать однородность двух текстовых коллекций путем сопоставления центроидов, характеризующих их терминологический состав. При этом к каждой из выборок необходимо применять одинаковые процедуры предварительной обработки: общий словарь стоп-слов, алгоритм стемминга (или лемматизации), способ расчета весов, программу машинного перевода (в случае использования переводных текстов) [3, 13].

Итак, у нас имеются две связанные выборки одинакового размера. Первая состоит из весов терминов в центроиде $C_1^{(i)}$, а вторая выборка содержит веса терминов центроида $C_2^{(i)}$.

Рассчитаем разности элементов связанных выборок:

$$Z^{(i)} = C_1^{(i)} - C_2^{(i)}.$$

Данные разности можно представить в виде неслучайной (систематической) и случайной составляющих [5]:

$$Z^{(i)} = \theta + e^{(i)},$$

где θ — неизвестный параметр, характеризующий неслучайные различия в терминологическом составе центроидов; $e^{(i)}$ — случайные величины, которые являются независимыми и принадлежат непрерывной совокупности, симметричной относительно нуля.

Проверяем нулевую гипотезу H_0 о том, что две исследуемые связанные выборки принадлежат однородной генеральной совокупности. Гипотеза H_0 означает, что изменение весов терминов в каждой из выборок носит случайный характер, альтернативная гипотеза H_1 предполагает, что терминологические составы центроидов не случайным образом («систематически») отличаются друг от друга. Эта гипотеза известна в литературе как «гипотеза сдвига» [1, 5, 20].

Если центроиды C_1 и C_2 обладают практически идентичным набором терминов с близкими значениями весов, то систематическая составляющая θ = 0, различия в разностях случайны ($Z^{(i)} = e^{(i)}$) и эффект значимого различия в терминологических составах отсутствует, что свидетельствует об однородности выборок.

В непараметрической статистике разработано несколько специализированных критериев для анализа связанных выборок [1, 6, 17, 20]. В данной работе используем критерий знаков и критерий знаковых рангов Вилкоксона (Wilcoxon Watched Pair Test).

Описание выборок, предварительная обработка текстовых данных и экспериментальная проверка однородности

В экспериментальных исследованиях используют три коллекции документов, полученных из различных источников. Под документом далее понимается библиографическое описание научной статьи (используются поля название, аннотация, ключевые слова).

Первая коллекция сформирована из русскоязычной цифровой библиотеки Elibrary [22]. Она состоит из документов по тематике «Интеллектуальный анализ данных». Для ее составления использован запрос: «Интеллектуальный анализ данных, Text Mining, Кластеризация, Классификация, Машинное обучение». Размер коллекции — 1000 документов.

К сожалению, Elibrary — практически единственный русскоязычный источник научной информации, из которого можно создавать датасеты по предметным областям. При этом доступное в Elibrary количество статей по тематике ИАД достаточно небольшое, что накладывает ограничения на размер обучающих тестовых выборок.

Остальные две коллекции по тематике ИАД формировались из широко известных англоязычных информационных ресурсов: электронного архива научных статей Корнелльского университета (arXiv. org) и электронной библиотеки ассоциации вычислительной техники ACM [22, 23] по запросу “Text Mining, Clustering, Classification, Machine Learning”. Размер каждой коллекции — 1000 документов. Полученные англоязычные публикации были переведены на русский язык с помощью Google-переводчика. В данной работе вопрос выбора наилучшей программы машинного перевода не рассматривался. Как представляется, при переводе с английского на русский язык большинство онлайн-переводчиков работает приблизительно одинаково, обеспечивая достаточно хорошее качество перевода.

Предварительная обработка текстовых данных для всех трех выборок включала применение словаря стоп-слов, лемматизацию и удаление редких слов, встретившихся в выборке менее пяти раз. Взвешивание терминов проводилось с помощью частоты терминов (при использовании *tf-idf*-взвешивания получены аналогичные результаты).

Нами исследованы различные длины центроидов, состоящие из 50, 100, 200, 500, 1000, 1500, 2000 терминов. Начиная с 1000 терминов размер центроида не влияет на проверку однородности (критерии показывают одинаковые результаты). Коэффициенты Жаккара, рассчитанные по формуле (1), для такой длины центроидов находятся в интервале [0,63; 0,69].

С помощью критерия знаков и критерия знаковых рангов Вилкоксона (для связанных выборок) на однородность были проверены следующие пары выборок:

ИАД (Elibrary) – ИАД (ACM);

ИАД (Elibrary) – ИАД (Arxiv);

ИАД (ACM) – ИАД (Arxiv).

При длине центроидов 50, 100 и 200 терминов гипотеза об однородности не подтверждается для всех выборок (при длине 500 на разных выборках получены противоречивые результаты — принимается H_0 или H_1). При этом наблюдаются невысокие значения коэффициентов Жаккара (менее 0,4). Начиная с длины центроидов в 1000 терминов для всех исследованных пар выборок по обоим критериям принимается гипотеза об однородности на уровне значимости α = 0,05.

Это позволяет сделать вывод, что исследуемые выборки, извлеченные по одной тематике как с русскоязычного ресурса, так и переведенные из англоязычных библиотек, являются однородными и представляется возможным их объединение в единую документальную коллекцию.

Заключение

Рассмотрена проблема оценки однородности выборок, формируемых из разных информационных ресурсов. Проанализированы причины возникновения сдвига данных при обработке и анализе документальных массивов по научным тематикам. Предложена процедура анализа терминологических различий в одноклассовых выборках на основе сопоставления их центроидов. Гипотеза об однородности проверена с помощью непараметрических критериев для связанных выборок (критерия знаков и критерия знаковых рангов Вилкоксона). Разработанная процедура использована для оценки однородности трех коллекций документов, полученных из различных (русско- и англоязычных) источников. Рассмотренный подход достаточно прост в использовании, позволяет принимать статистически обоснованные решения и может успешно применяться на практике для формирования крупных однородных документальных коллекций.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

ЛИТЕРАТУРА

1. **Орлов А. И.** Прикладная статистика. — М.: Экзамен, 2006. — 671 с.
2. **Бурков А.** Инженерия машинного обучения. — М.: ДМК Пресс, 2022. — 306 с.
3. **Мулатов Н. И., Мохов А. С., Толчеев В. О.** Способы построения текстовых коллекций для обучения классификаторов / Заводская лаборатория. Диагностика материалов. 2021. Т. 87. № 7. С. 76 – 84. DOI: 10.26896/1028-6861-2021-87-7-76-84
4. **Кафтанников И. Л., Парасич А. В.** Проблемы формирования обучающей выборки в задачах машинного обучения / Вестник ЮУрГУ. Серия Компьютерные технологии, управление, радиоэлектроника. 2016. Т. 16. № 3. С. 15 – 24.
5. **Холлендер М., Вульф Д.** Непараметрические методы статистики. — М.: Финансы и статистика, 1983 — 518 с.
6. **Орлов А. И.** Основные требования к математическим методам классификации / Заводская лаборатория. Диагностика материалов. 2020. Т. 86. № 11. С. 67 – 78. DOI: 10.26896/1028-6861-2020-86-11-67-78
7. **Lipton Z., Wang Y-X., Smola A.** Detecting and Correcting for Label Shift with Black Box Predictors / ArXiv: 1802.03916.2018.
8. Dataset Shift in Machine Learning / J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. Lawrence, Eds. — The MIT Press, 2022. — 248 p.
9. **Zhang K., Scholkopf B., Muandet K., Wang Z.** Domain Adaptation under Target and Conditional Shift / Proceedings of the 30th International Conference on Machine Learning. 2013. Vol. 28. N 3. P. 819 – 827.
10. **Subbaswamy A., Schulam P., Saria S.** Preventing Failures Due to Dataset Shift: Learning Predictive Models that Transport / Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. 2019. Vol. 89. P. 3118 – 3127.
11. **Parker B., Khan L.** Rapidly Labeling and Tracking Dynamically Evolving Concepts in Data Streams / IEEE 13th International Conference on Data Mining Workshops. 2013. P. 1161 – 1164.
12. **Ефимова И. В.** Формирование однородных обучающих выборок для задач медицинской диагностики / Труды 57-й Международной научной конференции МФТИ. 2014. С. 91 – 92.
13. **Evangeline M., Shyamala K.** Text Categorization Techniques: A Survey / International Conference on Innovative Practices in Technology and Management (ICIPTM). 2021. P. 137 – 142.
14. **Kreutz C. K., Schenkel R.** Scientific Paper Recommendation Systems: a Literature Review of recent Publications / ArXiv: 2201.00682.2022.
15. **Silambarasan M., Shathik J.** Ensemble Text Classifier: A Document Classification Technique to Predict and Categorizes Regularised and Novel Classes Using Incremental Learning / International Journal of Applied Engineering Research. 2017. Vol. 12. N 22. P. 12454 – 12459.
16. Understanding Dataset Shift and Potential Remedies. Technical Report. — Vector Institute, 2021. — 27 p.
17. **Орлов А. И.** Какие гипотезы можно проверять с помощью двухвыборочного критерия Вилкоксона / Заводская лаборатория. Диагностика материалов. 1999. Т. 65. № 1. С. 51 – 56.
18. **Орлов А. И.** Модель анализа совпадений при расчете непараметрических ранговых статистик / Заводская лаборатория. Диагностика материалов. 2017. Т. 83. № 11. С. 66 – 72. DOI: 10.26896/1028-6861-2017-83-11-66-72
19. **Орлов А. И.** Распределения реальных статистических данных не являются нормальными / Научный журнал КубГАУ. 2016. № 117. С. 71 – 90.
20. **Орлов А. И.** Методы проверки однородности связанных выборок / Заводская лаборатория. Диагностика материалов. 2004. Т. 70. № 7. С. 57 – 61.
21. **Frias-Blanco I., Campo-Avila J., Ramos-Jimenez G., Morales-Bueno R., Ortiz-Diaz A., Caballero-Mota Y.** Online and Non-Parametric Drift Detection Methods Based on Hoeffding's Bounds / IEEE Transactions on Knowledge and Data Engineering. 2014. Vol. 27. N 3. P. 810 – 823.
22. Digital Library Elibrary [cited February 3, 2023]. Available: https://eLibrary.ru
23. Electronic archive of scientific articles of Cornell University with open access [cited February 3, 2023]. Available: https://arxiv.org
24. Electronic Library of the Association for Computing Machinery ACM Digital Library [cited February 3, 2023]. Available: https://dl.acm.org

REFERENCES

1. **Orlov A. I.** Applied statistics. — Moscow: Ékzamen, 2006. — 671 p. [in Russian].
2. **Burkov A.** Machine Learning Engineering. — Moscow: DMK Press, 2022. — 306 p. [in Russian].
3. **Mulatov N. I., Mokhov A. S., Tolcheev V. O.** Methods of constructing text collections for training classifiers / Zavod. Lab. Diagn. Mater. 2021. Vol. 87. N 7. P. 76 – 84 [in Russian]. DOI: 10.26896/1028-6861-2021-87-7-76-84
4. **Kaftannikov I. L., Parasich A. V.** Problems of training sample formation in machine learning tasks / Vestn. UUrGu. Ser. Komp'yut. Tekhnol. Upr. Radioélektr. 2016. Vol. 16 N 3. P. 15 – 24 [in Russian].
5. **Hollender M., Wolf D.** Nonparametric methods of statistics. — Moscow: Finance and Statistics, 1983. — 518 p. [Russian translation].

6. **Orlov A. I.** Basic requirements for mathematical classification methods / Zavod. Lab. Diagn. Mater. 2020. Vol. 86. N 11. P. 67 – 78 [in Russian].
   DOI: 10.26896/1028-6861-2020-86-11-67-78
7. **Lipton Z., Wang Y-X., Smola A.** Detecting and Correcting for Label Shift with Black Box Predictors / ArXiv: 1802.03916.2018.
8. Dataset Shift in Machine Learning / J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. Lawrence, Eds. — The MIT Press, 2022. — 248 p.
9. **Zhang K., Scholkopf B., Muandet K., Wang Z.** Domain Adaptation under Target and Conditional Shift / Proceedings of the 30th International Conference on Machine Learning. 2013. Vol. 28. N 3. P. 819 – 827.
10. **Subbaswamy A., Schulam P., Saria S.** Preventing Failures Due to Dataset Shift: Learning Predictive Models that Transport / Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. 2019. Vol. 89. P. 3118 – 3127.
11. **Parker B., Khan L.** Rapidly Labeling and Tracking Dynamically Evolving Concepts in Data Streams / IEEE 13th International Conference on Data Mining Workshops. 2013. P. 1161 – 1164.
12. **Efimova I. V.** Formation of homogeneous training samples for medical diagnostics tasks / Proceedings of the 57th International Scientific Conference of MIPT. 2014. P. 91 – 92 [in Russian].
13. **Evangeline M., Shyamala K.** Text Categorization Techniques: A Survey / International Conference on Innovative Practices in Technology and Management (ICIPTM). 2021. P. 137 – 142.
14. **Kreutz C. K., Schenkel R.** Scientific Paper Recommendation Systems: a Literature Review of recent Publications / ArXiv: 2201.00682.2022.
15. **Silambarasan M., Shathik J.** Ensemble Text Classifier: A Document Classification Technique to Predict and Categorizes Regularised and Novel Classes Using Incremental Learning / International Journal of Applied Engineering Research. 2017. Vol. 12. N 22. P. 12454 – 12459.
16. Understanding Dataset Shift and Potential Remedies. Technical Report. — Vector Institute, 2021. — 27 p.
17. **Orlov A. I.** What hypotheses can be tested using the two-sample Wilcoxon criterion / Zavod. Lab. Diagn. Maters. 1999. Vol. 65. N 1. P. 51 – 56 [in Russian].
18. **Orlov A. I.** Model of coincidence analysis in the calculation of nonparametric rank statistics / Zavod. Lab. Diagn. Mater. 2017. Vol. 83. N. 11. P. 66 – 72 [in Russian].
    DOI: 10.26896/1028-6861-2017-83-11-66-72
19. **Orlov A. I.** Distributions of real statistical data are not normal / Scientific Journal of KubGAU. 2016. N. 117. P. 71 – 90 [in Russian].
20. **Orlov A. I.** Methods of checking the homogeneity of related samples / Zavod. Lab. Diagn. Mater. 2004. Vol. 70. N. 7. P. 57 – 61 [in Russian].
21. **Frias-Blanco I., Campo-Avila J., Ramos-Jimenez G., Morales-Bueno R., Ortiz-Diaz A., Caballero-Mota Y.** Online and Non-Parametric Drift Detection Methods Based on Hoeffding's Bounds / IEEE Transactions on Knowledge and Data Engineering. 2014. Vol. 27. N 3. P. 810 – 823.
22. Digital Library Elibrary [cited February 3, 2023]. Available: https://eLibrary.ru
23. Electronic archive of scientific articles of Cornell University with open access [cited February 3, 2023]. Available: https://arxiv.org
24. Electronic Library of the Association for Computing Machinery ACM Digital Library [cited February 3, 2023]. Available: https://dl.acm.org