УДК 681.3.06

К. Н. Шарутин, аспирант 1-го курса

А. С. Суркова, доктор техн. наук, профессор

Нижегородский государственный технический университет им. Р. Е. Алексеева

ОБЗОР МЕТОДОВ КЛАССИФИКАЦИИ ТЕКСТОВ

Аннотация: Целью данной работы является анализ различных методов классификации текстов, используемых на практике, их сильных и слабых сторон, а также повышение осведомленности о различных возможностях извлечения знаний в области интеллектуального анализа данных. Полученные результаты говорят о том, что необходимо изучить и понять природу данных, прежде чем приступать к их обработке. Кроме того, требуется повысить уровень автоматизации процессов классификации текстов в связи с растущим объемом данных и требуемой точностью. Также в ходе исследования было определено, что не существует окончательного алгоритма для конкретной задачи классификации текста с точки зрения автоматизации.
Ключевые слова: классификация текстов, машинное обучение, анализ, нейронные сети, статистические методы

Annotation: The purpose of this paper is to analyze the various methods of text classification used in practice, their strengths and weaknesses, as well as to raise awareness of the various possibilities for extracting knowledge in the field of data mining. The results suggest that it is necessary to study and understand the nature of melons before proceeding to their processing. In addition, there is a need to increase the level of automation of text classification processes due to the growing volume of data and the required accuracy. The study also determined that there is no definitive algorithm for a specific text classification task in terms of automation.
Key words: text classification, machine learning, analysis, neural networks, statistical methods

Введение

Неструктурированные данные остаются проблемой практически во всех областях применения интенсивного применения данных, таких как бизнес, университеты, исследовательские институты, государственные финансовые учреждения и технологические компании. Восемьдесят процентов данных о сущности (человеке, месте или вещи) доступны только в неструктурированной форме [1]. Они представлены в виде отчетов, электронной почты, просмотров, новостей и т. д.

Текстовая аналитика преобразует текст в числа, а числа, в свою очередь, привносят структуру в данные и помогают выявить закономерности. Чем более структурированы данные, тем лучше анализ и, в конечном счете, тем лучше будут приниматься решения. Также трудно обрабатывать каждый бит данных вручную и четко классифицировать их. Это привело к появлению интеллектуальных инструментов в обработке текста, в области обработки естественного языка, для анализа лексических и лингвистических паттернов [1].

Кластеризация, классификация и категоризация являются основными методами, используемыми в текстовой аналитике [1]. Это процесс присвоения, например, документу определенной классовой метки среди других доступных классовых меток. Таким образом, классификация текстов является обязательным этапом в раскрытии знаний. Целью данной статьи является анализ различных методов классификации текстов, используемых на практике, их распространения в различных прикладных областях, сильных и слабых сторон, а также современных тенденций исследований для повышения осведомленности о возможностях извлечения знаний.

Методология

Для данного исследования была изучена в общей сложности 41 исследовательская статья из баз данных, таких как IEEE, Science Direct, Springer, Google Scholar, Academia и других технических блогов, опубликованных в период с 2015 по 2020 год. Проблема управления большими потоками информации возникла в конце 2000-х годов, что увеличило потребность в процедурах обработки данных. Поэтому в этот период внимание исследователей к таким процедурам было повышенным.

Основные подходы к классификации текстов были далее организованы в виде древовидной структуры после анализа сходств и различий между этими различными подходами вместе с их соответствующими алгоритмами (рис. 1).

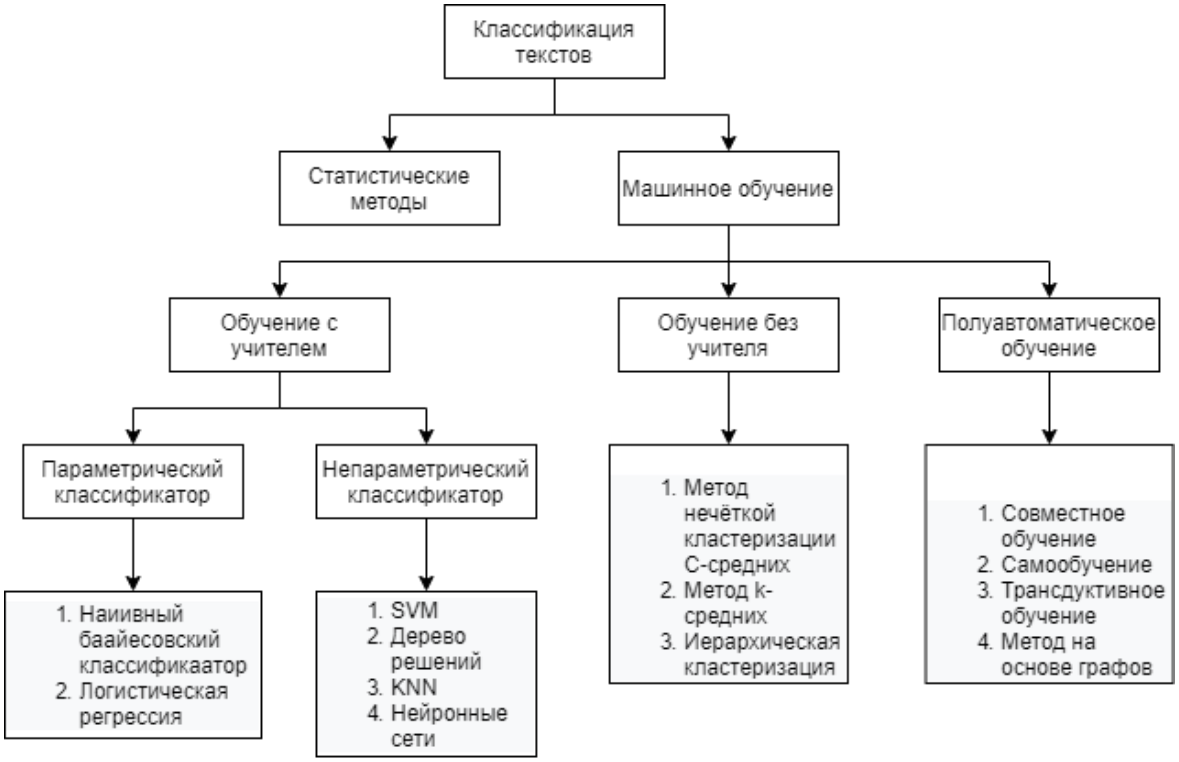

Рисунок 1 – Представление различных классификаторов

Использовались различные поисковые термины: текст + классификация, текст + классификация + алгоритмы, а также все подзаголовки, указанные на рисунке 1 в отношении классификации текста и машинного обучения. Области исследований были сгруппированы в соответствии с их более широкой областью, такой как статистические методы, алгоритмы обучения с учителем, без учителя и полуавтоматическое обучения.

Методы классификации текста

Алгоритмы классификации формируют основу методов интеллектуального анализа текста [1].

Как правило, метод классификации можно разделить на методы статистического и машинного обучения. Методы машинного обучения были специально изобретены для автоматизации [2]. На рисунке 1 алгоритмы машинного обучения разделены на с учителем, без учителя и полуавтоматические в соответствии с критериями обучения.

Ниже приведены некоторые методы классификации текстов и направления их исследования.

Статистический подход

Статистические методы — это чисто математические процессы, и они служат математической основой для всех других текстовых классификаторов. Они работают подобно компьютерной программе, выполняя инструкции без какой-либо собственной инициативы [2]. Для достижения хорошей классификации объем информации, обрабатываемой приложением, должен быть кратким, и это достигается за счет уменьшения размерности данных (количество переменных, которые должны учитываться в наборе данных, “возраст”, “пол”, “местность” и т. д., являются переменными). Данные могут быть достаточно сложны и многомерны. Методы извлечения статистических признаков, такие как Анализ главных компонент, Предвзятый дискриминантный анализ оказались лучшими

методами уменьшения размерности [2]. Они упорядочены по релевантности, но не подходят для нелинейных данных. Также они неэффективны для больших наборов данных. В будущем эти методы также могут быть использованы для классификации двоичного текста, чтобы определить, является ли конкретное электронное письмо спамом или нет.

Подход с использованием машинного обучения

Увеличение объема, скорости и разнообразия данных потребовало автоматизации методов обработки текста, включая классификацию текста. В некоторых ситуациях определение набора логических правил с использованием методов инженерии знаний и на основе экспертных заключений для классификации документов помогает автоматизировать задачу классификации. Классификацию текста можно разделить на три категории: классификация с учителем, классификация без учителя и полуавтоматическая классификация текста [4].

В терминологии машинного обучения проблема классификации подпадает под принцип обучения с учителем, когда система обучается и тестируется на знаниях о классах до фактического процесса классификации. Обучение происходит, когда помеченные данные недоступны. Этот процесс сложен, а также ресурсозатратен. Он подходит для больших данных. Полуавтоматическое обучение происходит, когда данные частично размечены и частично не размечены [10].

Использование большего набора данных позволяет добиться меньших ошибок при классификации. Известно также, что выбор подходящих алгоритмов для конкретного набора данных играет важную роль в классификации текста.

Обучение с учителем

Обучение с учителем - самое дорогое и крайне сложное из представленных. Данное понятие подразумевает вмешательство человека при присвоении меток классам, что невозможно в больших наборах данных. Обучение с учителем становится дорогостоящим, когда существуют различные распределения данных, различные выходные данные и различные пространства признаков. Подобный вид обучения принято делить на два подхода: с применением параметрического классификатора и непараметрического.

Параметрические классификаторы

Классификатор, который может суммировать данные на основе базовых параметров, называется параметрическим классификатором [4]. Логистическая регрессия и Наивные байесовские алгоритмы являются параметрическими классификаторами.

Наивный байесовский классификатор

Это вероятностный классификатор, часто используемый в машинном обучении, но также может использоваться как статистический метод. Он в основном используется при предварительной обработке данных из-за простоты вычислений. Байесовское рассуждение и вероятностный вывод используются для предсказания целевого класса. Атрибуты играют важную роль в классификации, поэтому присвоение атрибутам различных весовых значений потенциально может улучшить производительность [11].

Эффективность наивного байесовского классификатора зависит от точности оцениваемых условных вероятностей. Трудно точно оценить эти условия при недостатке обучающих данных. Поэтому некоторые методы метаэвристики, такие как генетические алгоритмы, дифференциальная эволюция, используются для оценки условных вероятностей. В некоторых случаях преимущества данного классификатора оспариваются условным допущением независимости между атрибутами, которое имеет свое влияние на эффективность классификации [11]. Для улучшения эффективности используются различные методы мета-обучения, такие как расширение структуры, выбор атрибутов, преобразование частоты, взвешивание атрибутов, взвешивание экземпляров и локальное обучение. Таким образом, наивные байесовские классификаторы просты в реализации и одновременно эффективны с точки зрения степени достоверности. Эти возможности делают классификатор подходящим для решения проблем обработки естественного языка.

Логистическая регрессия

Логистическая регрессия в обучении с учителем представляет собой выбор лучших параметров для маркировки для достижения эффективных результатов классификации. Логистическая регрессия похожа на линейную тем, что в ней тоже требуется найти значения коэффициентов для входных переменных. Разница заключается в том, что выходное значение преобразуется с помощью нелинейной или логистической функции.

Логистическая функция выглядит как большая буква S и преобразовывает любое значение в число в пределах от 0 до 1. Это весьма полезно, так как мы можем применить правило к выходу логистической функции для привязки к 0 и 1 (например, если результат функции меньше 0.5, то на выходе получаем 1) и предсказания класса.

Благодаря тому, как обучается модель, предсказания логистической регрессии можно использовать для отображения вероятности принадлежности образца к классу 0 или 1. Это полезно в тех случаях, когда нужно иметь больше обоснований для прогнозирования.

Как и в случае с линейной регрессией, логистическая регрессия выполняет свою задачу лучше, если убрать лишние и похожие переменные. Модель логистической регрессии быстро обучается и хорошо подходит для задач бинарной классификации.

Непараметрический классификатор

Классификатор, который не может суммировать данные на основе базовых параметров, называется непараметрическим классификатором. Машина опорных векторов, алгоритм *k*-ближайшего соседа, деревья решений и нейронные сети являются непараметрическими классификаторами.

Метод опорных векторов

Метод опорных векторов (*SVM*) является одним из применимых алгоритмов машинного обучения, который используется для различных задач классификации [7]. Чаще всего применяется в анализе кредитных рисков, медицинской диагностике, классификации текстов и извлечении информации. *SVM* особенно подходят для данных большого размера.

Вариации стандартных ядер, известные как кастомизированные ядра, повышают производительность алгоритмов, поскольку включают в себя фоновые детали для категоризации текста. Одним из таких кастомизированных ядер является ядро значения класса, которое используется для сглаживания терминов документов с использованием значений терминов на основе классов [7]. Одно и то же ядро может использоваться для получения неявной семантической информации при вычислении сходства между документами в будущем.

Также метод опорных векторов предлагается использовать при полуавтоматической кластеризации для классификации текстов. Это помогает определить категорию текста из нескольких компонентов. В данном случае размеченные данные используются для повышения производительности системы, поскольку они способствуют эффективной оценки параметров.

Деревья решений

Деревья решений работают в определенной последовательности, чтобы проверить решение на соответствие определенному пороговому значению среди доступных значений. Тестирование происходит по определенным логическим правилам, аналогичным весам в нейронных сетях. На этапе роста дерева обучающий набор разбивается на части, а на этапе обрезки обобщаются данные по нему. Деревья, основанные на ансамбле, используют методы бустинга и бэггинга для объединения более чем одного классификатора, который использует различные правила принятия решений для различных наборов данных [8]. Данные ансамбли показали замечательную производительность по сравнению с обычными деревьями решений, однако вычислительные затраты увеличиваются по мере добавления входных запросов [8].

Деревья решений обладают низкой эффективностью при обработке многомерных данных. Для решения этой проблемы предлагаются кластерные деревья. Инкрементные

деревья принятия решений лучше всего подходят для потоков данных, поскольку они обладают способностью стабилизироваться в соответствии с накапливающимися данными. Они использует несколько атрибутов для обучаемых функций.

Кроме того, еще одной областью, требующей широкого изучения, является область оптимизации деревьев решений.

Алгоритм k-ближайших соседей

Алгоритм k-ближайшего соседа (k-NN) работает по принципу ближайших обучающих выборок, те точки данных, которые близки друг к другу, принадлежат к одному определенному классу [6]. Большую сложность представляет определение значения k, кроме того, вычислительная сложность также возрастает с увеличением размерности. Для снижения затрат на вычисление значения k используется древовидный k-NN. Данный алгоритм позволяет уменьшить область поиска за счет более совершенных методов обхода [6]. Также среди существующих подходов можно выделить подход с использованием повторной выборки, чувствительный к затратам на обучение. Алгоритм k-NN чаще всего применим для классификации экземпляров на основе контекста точек данных путем голосования большинством голосов. Этот метод подходит только для небольших наборов данных.

Искусственные нейронные сети

Искусственные нейронные сети (ANN) имитируют работу человеческого мозг при принятии решения. Они работают, обучаются и эволюционируют с минимальным вмешательством человека или вообще без него. Для классификации данных предлагается конкурентный алгоритм коэволюции, основанный на нейросетевой модели. Радиальная базисная функция является компонентом ANN, поскольку в ней используются более быстрые алгоритмы обучения. Данная функция имеет компактную сетевую архитектуру, которая повышает точность классификации. Кроме того, эволюционирующие алгоритмы имеют тенденцию хорошо работать в динамических средах, приспосабливаясь на лету и адаптируясь к "нечетким" характеристикам [9].

Нейронные сети также применяются в случае, когда требуется иерархический подход к классификации с несколькими метками. Данный вид классификации сложен, поскольку каждая выборка может принадлежать более чем к одному классу, и прогнозы одного уровня подаются в качестве входных данных на следующий уровень для принятия окончательного решения [10].

ANN имеют хорошую прикладную ценность, потенциал развития, а также нет необходимости обучать отдельные бинарные классификаторы для многоклассовых задач, поэтому они формируют лучшие базовые классификаторы в ансамблевом подходе.

Обучение без учителя

Обучение без учителя — это алгоритм машинного обучения, при котором испытуемая система спонтанно обучается выполнять поставленную задачу без вмешательства со стороны экспериментатора. Изначально данный тип алгоритмов кажется сложным, но, когда в модель поступает больше данных, алгоритм совершенствуется. Анализ основных компонентов, кластеризация и самоорганизующиеся метки часто используются в обучении без учителя. Во многих случаях экспертные знания, необходимые для маркировки образцов, либо отсутствуют, либо их недостаточно. В этом случае самоорганизующиеся карты Кохонена и коэффициент корреляции используются для кластеризации документов и использования их для маркировки документов для дальнейшей классификации [12]. Данный способ позволяет избавиться от проклятия размерности и экспертного вмешательства. Этот вид гибридной модели больше подходит для больших объемов данных.

Некоторые из известных алгоритмов обучения без учителя — это обнаружение аномалий, обучение Хебба, алгоритм максимизации ожиданий, анализ главных компонент, независимый компонентный анализ, неотрицательная матричная факторизация, сингулярная декомпозиция.

Полуавтоматическое обучение

Полуавтоматическое обучение — это сочетание методов обучения с учителем и без. Этот тип обучения использует небольшое количество маркированных данных и большое количество немаркированных данных для обучения. Метки назначаются путем объединения помеченных и немеченых экземпляров, поскольку немеченые данные смягчают влияние недостаточного количества помеченных данных на точность классификатора. Некоторые из методов полуавтоматического обучения включают в себя самоподготовку, самообучении, совместное обучение, трансдуктивный метод опорных векторов, генеративные модели и графовые методы [12].

Традиционные подходы к классификации текста становятся бесполезными в случае отсутствия помеченных данных для определенного класса набора данных, например, помеченные данные доступны только для положительных выборок, а не для отрицательных. Для этого же рекомендуется полуавтоматический ансамблевый алгоритм. Недоступный класс извлекается приблизительно из набора данных и устанавливается в качестве помеченного образца.

Ансамблевый классификатор итеративно строит границу между положительными и отрицательными классами для дальнейшей аппроксимации отрицательных данных, поскольку отрицательные данные смешиваются с положительными данными. Таким образом, без необходимости в обучающих выборках классификация достигается с помощью гибридного подхода. Это исключает затраты на ручную маркировку данных, особенно больших данных.

Применение полу управляемых алгоритмов очень полезно в требованиях к фильтрации информации. Роль полуавтоматических алгоритмов в иерархической классификации с несколькими метками — это область, которая требует более глубокого исследовании.

Полученные данные

На основе исследования, проведенного для данной статьи, было установлено, что наиболее широко используемые методы классификации текстов соответствуют полуавтоматическому подходу обучения [10]. Поскольку данный подход обладает потенциалом для повышения эффективности классификации за счет комбинированных преимуществ методов обучения с учителем и без. Установлено, что метод активного обучения используется для снижения временных затрат, связанных с выбором только наиболее подходящего экземпляра для классификации выборки (итеративное обучение с учителем).

Производительность классификатора зависит от характера анализируемых данных, а также от типа хранилища данных. Хранение данных является решающим этапом в поддержании больших наборов данных и доступе к ним для обнаружения признаков. Используются две формы, такие как пакетная обработка и потоковое поглощение. Использование облачных технологий, таких как Amazon Redshift, Google BigQuery и Snowflake, способствует масштабируемости хранилищ данных за счет возможности облачной поддержки.

Данное исследование показало, что извлечение признаков, семантическая обработка, эффективность алгоритмов, гетерогенные данные, автоматизация аудита, масштабируемость данных, нарушение данных и принятие решений в реальном времени — вот некоторые из областей, которые требуют дальнейших исследований и разработок в отношении процедур классификации текстов.

Кроме того, можно утверждать, что не существует окончательного алгоритма для конкретной задачи классификации текста с точки зрения автоматизации.

Различные алгоритмы, рассмотренные в этом исследовании, суммированы в соответствии с их сильными и слабыми сторонами в Таблице 1.

Таблица 1 – Краткое изложение различных методов классификации текстов

Метод	Преимущества	Недостатки	Применение
Логистическая регрессия	Простая оценка параметров, хорошо работает для категориальных прогнозов.	Требуется большой размер выборки, не подходит для нелинейных задач	Финансовое прогнозирование, прогнозирование стоимости программного обеспечения, обеспечение качества программного обеспечения, Интеллектуальный анализ данных о преступлениях
Наивный байесовский классификатор	Быстрый классификатор, требует меньшего времени обучения, применяется как для бинарных, так и для много классовых задач	Отсутствие взаимодействия между признаками Вычисленные вероятности являются не математически точными, а относительными.	Фильтрация электронной почты, классификация статей на основе контента, анализ настроения/эмоций.
SVM	Параметр регуляризации позволяет избежать дополнительной подгонки параметров. Использование ядра включает дополнительные значения.	Выбор наилучшего ядра, а также время, затраченное на обучение и тестирование.	Хорошо подходит для биологических наборов данных, гипертекстовой категоризации и т. д.,
Дерево решений	Простота интерпретации, легко визуализировать, быстрота обучения и прогнозирования	Чувствительность к шумам входных данных, сложен для неопределенных и многозначных атрибутов.	Маркетинговые данные, информация о клиентах
Метод k-ближайших соседей	Более простая реализация, гибкий выбор функций, хорошо подходит для много классовых задачи	Поиск ближайших соседей и оценка оптимального значения k	Рекомендательные системы, медицина
Нейронные сети	Простота использования, скорость реализации, приближены по возможностям к любым предыдущим алгоритмам	Требует больших обучающих и тестовых данных, большая часть операций скрыта и труднодоступна для повышения точности.	Прогноз продаж, валидация данных, управление рисками и целевой маркетинг

Метод	Преимущества	Недостатки	Применение
Обобщенный алгоритм Хебба	Подходит для многоклассовых моделей в нейронных сетях. Легко интерпретировать послойные операции.	Возможность принимать только ортогональные входы, которые не являются коррелированными.	Подходит для распознавания изображений и речи в моделях искусственного интеллекта.
Выявление аномалий	Мало затратен по памяти, отсутствие параметров подбора, меньшая алгоритмическая сложность	Сложность в выявлении правил, иногда возникают выбросы, почти похожие на исходные паттерны.	Обнаружение мошенничества, отчетность о неисправностях, медицина
EM-алгоритм	Лучше подходит для гетерогенных наборов данных и прост в реализации, устойчив к шумам	При неудачной инициализации сходимость алгоритма может оказаться медленной	Реконструкция изображений, вероятностные контекстно-свободные грамматики и управление рисками в теории реагирования на элементы.
Сингулярное разложение	Устойчив к численным ошибкам, снижает размерность данных	Данные должны быть детрендированы перед применением алгоритма, и они должны содержать выбросы или аномалии.	Приложения для обработки цифровых сигналов и изображений. Рекомендательные системы для прогнозирования рейтингов.

Заключение

В ходе выполнения данной работы были изучены различные методы классификации текстов с их сильными сторонами, возможностями и слабыми сторонами в извлечении признаков из данных. На этом этапе крайне важно осознать проблемы, существующие в методах классификации текстов, чтобы легче было судить о различных классификаторах.

После изучения существующих методов классификации можно сделать вывод, что полуавтоматическая классификация текстов приобретает все большее значение благодаря своей эффективности. Такой подход позволяет снизить временные затраты. Среди открытых вопросов по-прежнему остаются такие вопросы, как повышение производительности, обработка больших данных, выбор функций, дисбаланс данных.

Также можно сделать вывод, что по-прежнему нецелесообразно использовать один конкретный классификатор для конкретной задачи. Тем не менее, количество "проб и ошибок" для выбора наилучшего классификатора можно было бы минимизировать на основе информации, представленной в этом исследовании.

Библиографический список:

1. Aliwy A. H., Ameer E. Comparative study of five text classification algorithms with their improve-ments. // International Journal of Applied Engineering Research. 2017. №12. С. 4309-4319.
2. Benites F., Sapozhnikova E. Improving scalability of ART neural networks. // *Neurocomputing*. 2017. №*230*. С. 219–229. URL:https://doi.org/10.1016/j.neucom.2016.12.022 (дата обращения: 21.03.2021).
3. Khan A., Baharudin B., Lee L. H., Khan K. A review of machine learning algorithms for text-documents classification. *// Journal of Advances in Information Technology*. ,2010. №1. С. 4-20.
4. Gomez J. C., Boiy E., Moens M. F. Highly discriminative statistical features for email classification. // *Knowledge Information Systems*. 2012. №31. С. 23–53. URL:https://doi.org/10.1007/s10115-011-0403-7 (дата обращения: 15.02.2021).
5. Asadi S., Shahrabi J. ACORI: A novel ACO algorithm for rule induction. // *Knowledge-Based Systems*. 2016. №97. С. 174-187.
6. Maillo J., Ramfrez S., Triguero I., Herrera F. kNN-IS: An iterative spark-based design of the k-nearest neighbors' classifier for big data. // *Knowledge-Based Systems*. 2018. №117. С. 3-15.
7. Mirzamomen Z., Kangavari M.R. Evolving fuzzy min–max neural network-based decision trees for data stream classification. // *Neural Processing Letters*. №45(1). 2017. С. 341–363. URL:https://doi.org/10.1007/s11063-016-9528-8 (дата обращения: 15.12.2020).
8. Nidhi Gupta V. Recent trends in text classification techniques. // *International Journal of Computer Applications*. 2020. №35(6). С. 45-51.
9. Rubin T. N., Chambers A., Smyth P., Steyvers,M. Statistical topic models for multi-label document classification. // *Machine Learning*, 2016. №88. С. 157–208.
10. Stas J., Juhar J., Hladek D. Classification of heterogeneous text data for robust domain-specific lan-guage modeling. // *EURASIP Journal on Audio, Speech, and Music Processing*. 2016. №1. С. 1-12.
11. Tsangaratos P., Ilia I. Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of model's complexity and training dataset size. // *Catena*. 2017. №145. С. 164–179. URL:https://doi.org/10.1016/j.catena.2016.06.004 (дата обращения: 03.01.2021).
12. Sun Z., Ye Y., Deng W., Huang Z. A cluster tree method for text categorization. // Procedia Engineering, 2017. №15, С. 3785-3790.