УДК 004.827
DOI 10.17513/snt.39908

# АДАПТИВНАЯ ТЕХНОЛОГИЯ КЛАССИФИКАЦИИ С ОБРАТНОЙ СВЯЗЬЮ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

**Митин Г.В., Панов А.В.**

*Московский государственный университет информационных технологий, радиотехники и электроники (МИРЭА), Москва, e-mail: grigory.mitin@mail.ru, insegnantenew@yandex.ru*

В данной работе поднимается проблема ограниченной применимости методов классификации из области машинного обучения в условиях наличия неопределенности организации входных данных. Особенно остро эта проблема проявляется при работе с потоковыми данными, вследствие невозможности предварительной оценки данных для обучения классификатора. Для преодоления означенных проблем необходимо применить технологию, способную подстраиваться под изменяющийся набор данных и производить дообучение классификатора с сохранением показателей качества и преемственности по эволюции классов. Описанная в статье реализация технологии адаптивной классификации построена на базе расширяемой цепочки бинарных классификаторов, способных к дообучению. Особые механизмы сбора статистики совпадений и неопознанных данных в процессе классификации позволяют организовать процесс дообучения составного классификатора. Особенностями данного подхода являются возможность автоматизации дообучения и отсутствие потерь неопознанных данных. Наличие нескольких петель обратной связи между логическими блоками обеспечивает непрерывный контроль качества классификации и эффективное предотвращение таких проблем, как переобучение либо недообучение классификатора. Описанная технология позволяет не только работу с потоковыми данными, но и практически полную автоматизацию процесса эволюции классификатора, динамически подстраивающегося под постоянно изменяющиеся входные данные. Подобная инновационная технология имеет целый ряд преимуществ по отношению к распространенным подходам к применению методов классификации из области машинного обучения.

**Ключевые слова: машинное обучение, обучение с учителем, адаптивная классификация**

# ADAPTIVE CLASSIFICATION TECHNOLOGY WITH FEEDBACK, BASED ON MACHINE LEARNING METHODS

**Mitin G.V., Panov A.V.**

*Moscow State University of Information Technologies, Radio Engineering and Electronics (MIREA), Moscow, e-mail: grigory.mitin@mail.ru, insegnantenew@yandex.ru*

This article raises the problem of limited applicability of classification methods based on machine learning technologies in case of uncertainty in data organization. This problem is critical for streaming data processing, due to impossibility of preliminary data analysis for classifier training. To overcome these problems, it is necessary to apply a technology that can adapt to a changing data set and retrain the classifier while maintaining quality and acceptance indicators for the evolution of classes. The implementation of adaptive classification technology described in the article is based on an extensible chain of binary classifiers capable of further training. Special mechanism for collecting statistics of coincidences and unidentified data in the classification process allows to organize retraining process of the composite classifier. The features of this approach are the possibility of automating additional training and the absence of loss of unidentified data. The presence of several feedback loops between logical blocks ensures continuous quality control of classification and effective prevention of problems such as over-training or under-training of the classifier. The described technology allows not only working with streaming data, but also almost complete automation of the classifier evolution process, dynamically adjusting to constantly changing input data. Such an innovative technology has a number of advantages in relation to common approaches to the application of classification methods from the field of machine learning.

**Keywords: machine learning, supervised learning, adaptive classification**

Все большее место в нашей жизни занимают технологии машинного обучения [1; 2]. Они уже используются для анализа предпочтений пользователей интернет-магазинов и социальных сетей, а также для предсказания поведения цен на бирже [2].

Возможности машинного обучения поражают. Уже существуют мощные нейросети, способные по текстовым описаниям создавать музыку, картины и даже писать программный код простых приложений.

Однако на практике использование методов машинного обучения сопряжено с рядом трудностей. Каждая прикладная задача требует выбора подходящего метода и его адаптации к специфике предметной области. Такая адаптация часто требует отдельного анализа обрабатываемых данных с целью определения их ключевых особенностей в рамках задачи, однако проведение такого анализа не всегда возможно.

Одними из наиболее распространенных методов машинного обучения стали методы классификации. Их работа состоит в распределении информационных объектов на группы – классы. Правила такого разделения задаются таблично, в виде обучающей выборки – набора примеров такого разделения.

Цель исследования – создание технологии адаптивной классификации с помощью методов машинного обучения, способной работать с потоковыми данными и автоматически реагировать на изменения входных данных с донастройкой решающего алгоритма и обратной связью по качеству классификации и необходимости дообучения.

**Материалы и методы исследования**

Материалы и методы исследования – при проведении исследования использованы анализ предметной области, анализ распространенных проблем методов обучения с учителем, моделирование процесса обучения классификатора.

*Ограничения методов классификации*

Любая неопределенность входных данных серьезно ограничивает применимость методов классификации в чистом виде. Одним из ограничивающих факторов является необходимость наличия обучающей выборки до начала работы с реальными данными. При этом качество обучающей выборки напрямую влияет на качество конечной классификации. Обеспечение качества обучающей выборки требует предварительной оценки входных данных, что трудно осуществить на практике. При работе с потоковыми данными такая оценка практически невозможна, так как данные поступают неограниченно и, таким образом, делают любые оценки постоянно устаревающими.

Простое расширение обучающей выборки при поступлении новых данных не решает проблему, так как ведет к переобучению классификатора. Следовательно, необходим нетривиальный подход, который позволит не только обновить, но и сбалансировать обучающую выборку.

*Выбор пути преодоления ограничений*

Для решения проблемы обеспечения качества обучающей выборки необходим подход, собирающий информацию о распределении данных в процессе обработки. Это предполагает изменение обучающей выборки в соответствии с собранной информацией, дообучение классификатора и контроль качества классификации.

Первоначальная оценка числа классов и распределения информационных объектов в потоке данных может оказаться неверной. В связи с этим критически важно определять информационные объекты, не принадлежащие существующим классам, и быть в состоянии продолжить анализ без потери данных о них.


*Рис. 1. Общая схема адаптивной классификации*

Предлагаемый адаптивный подход к классификации данных предполагает наличие обратной связи от механизма, обеспечивающего автоматический контроль качества классификации (рис. 1). Контроль качества делает возможным корректное и своевременное внесение изменений в процесс классификации.

Ключом к поддержанию высокого качества классификации в потоке данных является своевременное дообучение. Информационные объекты, не опознанные как элементы существующих классов, могут быть как элементами нового, ранее неизвестного класса, так и элементами одного из известных классов, потерянных вследствие недостатка информации о реальном распределении данных. Оба случая говорят о том, что классификатор начинает устаревать и требует обновления. Кроме того, важно не допустить неконтролируемый рост обучающей выборки, так как это приведет к переобучению.

Обновление и дообучение классификатора тесно связано с обновлением обучающей выборки. Такие процессы, как сбор данных для разметки и контроль качества полученной выборки, стоит доверить алгоритмам, в то время как окончательное решение по разметке данных должен принимать человек.

На рис. 1 можно видеть две петли обратной связи. Одна из них связывает контроль качества с механизмом обучения, поддерживая качество обучения на должном уровне. Вторая соединяет механизм обучения и классификатор и позволяет автоматически обновлять классификатор во время работы.

Существующие методы классификации не поддерживают подобную функциональность.

### *Реализация ключевых механизмов адаптивной классификации с обратной связью*

Рассмотрим ситуацию, при которой информационные объекты в потоке данных представлены векторами признаков. Подобные векторы также называют точками данных, по аналогии с материальными точками. Именно в таком формате принимают данные большинство методов машинного обучения.

Также предположим, что перед началом классификации данные проходят подготовку, включающую в себя компоновку, нормализацию и очистку от информационного шума, как показано на рис. 3. Подобные задачи могут взять на себя методы кластеризации, нацеленные на работу в условиях информационного шума [3]. Им не требуется обучающая выборка, что делает их эффективно функционирующими в условиях не определенной заранее организации входных данных [4]. Подготовка данных с использованием методов кластеризации гарантирует, что точки данных образуют несколько групп. Метки кластеров и другая информация, полученная в процессе подготовки данных, не участвует в классификации напрямую, однако используется для облегчения работы человеку, ответственному за финальный этап разметки неопознанных данных.

Механизм работы большинства методов классификации не подразумевает возможности появления точек данных, не принадлежащих ни одному классу [5]. Для преодоления этого ограничения следует воспользоваться составным классификатором, функциональная схема которого изображена на рис. 2. Такой классификатор состоит из блоков, реализующих один из методов классификации в режиме бинарной классификации, то есть разделения входящих точек на «принадлежащие» своему классу и «не принадлежащие». Одним из кандидатов на роль такого метода является «Naive Bayes», показывающий хорошие результаты при относительно небольшой обучающей выборке и имеющий весьма ограниченный набор гиперпараметров для настройки [6].

Точка данных поступает на вход первому блоку классификации. Если он опознает ее как принадлежащую своему классу, точка получает метку соответствующего класса. Если точка не была опознана как точка этого класса, она передается следующему в цепочке блоку классификации. Процесс продолжается до тех пор, пока точка не получит метку класса или пока не будут пройдены все блоки. Точки, не получившие метку класса, добавляются в список неопознанных точек данных для дальнейшего изучения.

Таким образом, классы, которым соответствуют блоки классификации, находящиеся раньше по цепочке, имеют больший приоритет, чем те, что стоят после них. Важно уточнить, что метка класса и его приоритет никак не связаны между собой. Класс с меткой «1» может быть соотнесен как с первым и самым приоритетным блоком, так и с одним из блоков в середине или в конце цепочки.

По достижении определенного числа элементов список неопознанных точек данных отправляется на разметку для обновления обучающей выборки, как показано на рис. 3.

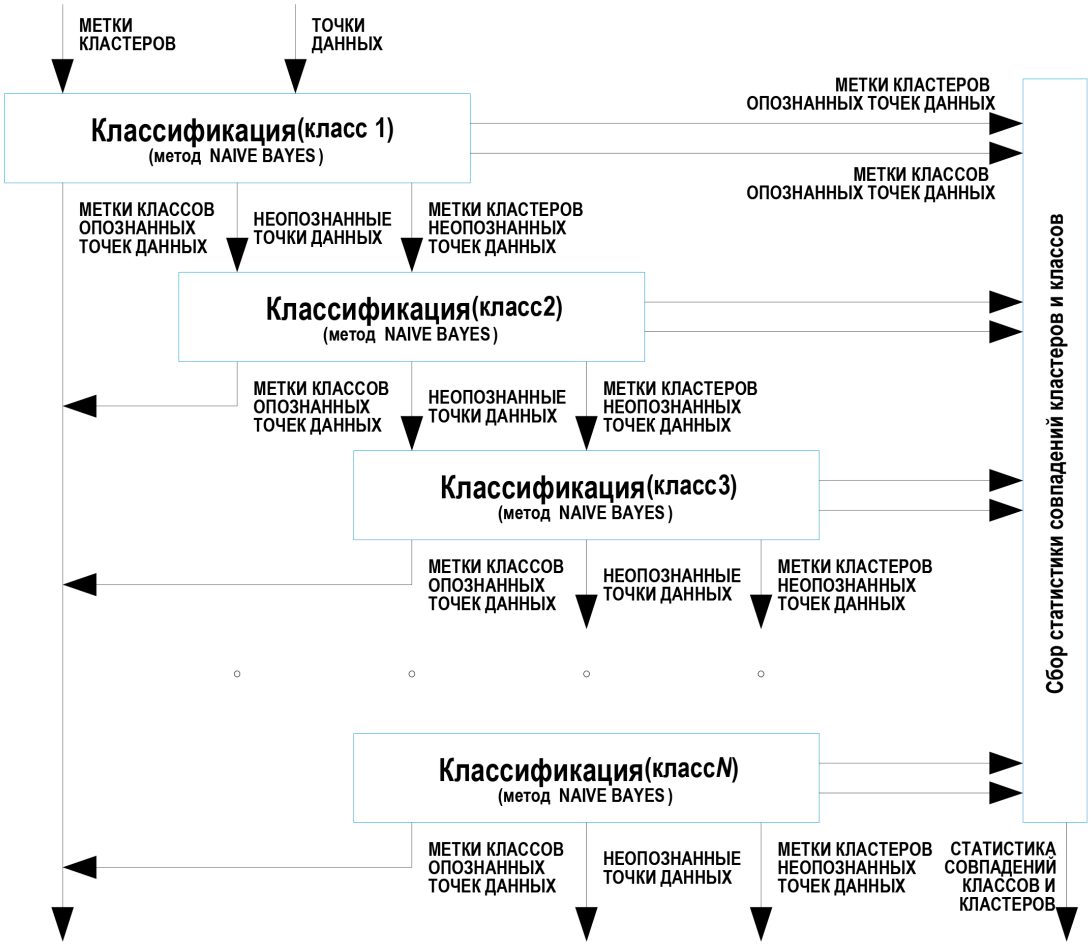


*Рис. 2. Функциональная схема составного классификатора со сбором статистики*

По отдельности точки данных малоинформативны. Человеку гораздо легче ориентироваться в особенностях распределения данных, когда он видит группу точек, в частности повышается вероятность обнаружения нового класса.

Также на разметку передается информация, полученная при подготовке данных, и статистика, собранная составным классификатором (рис. 3). Эти данные не только используются в качестве справочного материала для человека, но и позволяют провести предварительную разметку точек с высокой вероятностью попадания в тот или иной класс.

Размеченные точки данных добавляются в общий список размеченных точек, называемый учебным набором. Впоследствии учебный набор используется для обучения. Каждый раз, когда размер набора изменяется достаточно сильно, он проходит контроль качества по параметрам непротиворечивости и достаточного количества элементов каждого класса. Только при достижении минимальных требований по обоим параметрам набор можно использовать.

Противоречивость учебного набора относительно отдельного класса выражается отношением числа одинаковых или очень похожих точек, принадлежащих к разным классам, к числу элементов этого класса. Чем больше этот параметр, тем сложнее классификатору отделить этот класс от других и тем чаще будут возникать конфликты между блоками классификации. Этот параметр не должен превышать порогового значения.

Параметр достаточного количества элементов призван бороться с проблемой недообучения и дисбалансом классов. Он одинаков для каждого класса и задается вручную, так как зависит от особенностей задачи и метода классификации.

Если качество учебного набора оказалось недостаточным, механизм подготовки обучающей выборки отправляет запрос получения новых данных по обратной связи с механизмом сбора неопознанных данных (рис. 3).

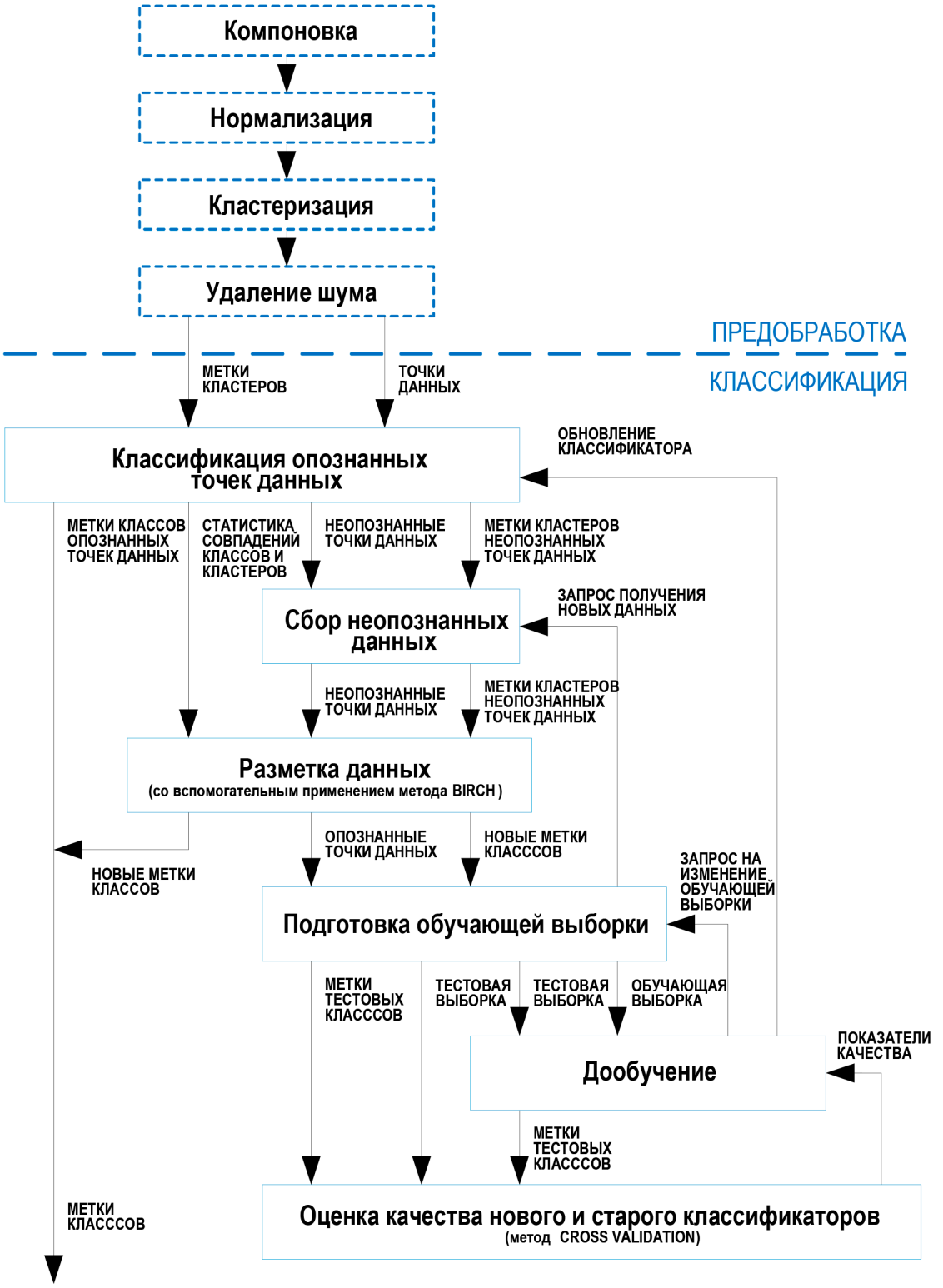


*Рис. 3. Общая функциональная схема полного цикла адаптивной классификации*

В ответ на него список неопознанных точек данных будет передавать данные на разметку вне очереди, пока учебный набор не достигнет требуемого качества. Этот шаг опирается на предположение, что увеличение числа точек в учебном наборе дает более точное представление о распределении данных в потоке, и в частности о каждом классе.

Достаточно сильные изменения учебного набора говорят о необходимости обновления составного классификатора. Из данных учебного набора генерируются обучающая выборка и тестовая выборка, каждая из которых содержит элементы всех классов. Обучающая выборка используется для обучения составного классификатора и первичной проверки качества классификации. Лишь после того, как все блоки классификации покажут приемлемый результат на обучающей выборке, они допускаются до дальнейших тестов. Обучающая выборка проходит проверку, аналогичную оценке качества учебного набора. Тестовая выборка, в свою очередь, используется для итоговой проверки качества классификации. Для чистоты оценки она не должна содержать элементов обучающей выборки.

Если учебный набор изменяется не первый раз, то тестовая и обучающая выборка обязательно содержат часть точек с предыдущей итерации. Таким образом исключается возможность противоречий между старым и новым составными классификаторами – возникает преемственность. В остальном обучающая выборка генерируется «с нуля», позволяя «перемешивать» выборку, оказавшуюся неудачной. Стоит упомянуть, что выборка ограничена как минимально достаточным, так и максимальным числом элементов каждого класса, чтобы избежать дисбаланса классов, а также проблем переобучения и недообучения. Подробный обзор проблем при формировании обучающей выборки представлен в [7].

На тестовом наборе проходит проверку как новый классификатор, сгенерированный на основе данных из обучающей выборки, так и соответствующий набор блоков действующего классификатора. Блоки нового классификатора, показавшие результат лучший, чем соответствующие блоки действующего классификатора, используются для обновления классификатора, в то время как блоки, показавшие неудовлетворительный результат, проходят повторное обучение, при этом меняется состав обучающей выборки для соответствующих классов. Обновление классификатора по блокам способствует повышению скорости обновления без потери качества.

Обратная связь по контролю качества, изображенная на рис. 1 и 3, реализует оперативное изменение обучающей выборки. После составления списка классов, показавших неудовлетворительные результаты на тестовой выборке, этот список передается механизму подготовки обучающей выборки в составе запроса на изменение обучающей выборки. Ответом на запрос является новая обучающая выборка, в достаточной мере отличная от предыдущей. После этого процесс обучения и проверки качества классификатора начинается заново, с использованием новой обучающей выборки. Заметим, что во многих существующих системах переработкой обучающей выборки занимается человек, в то время как описанная технология позволяет автоматизировать ее.

Обновление действующего классификатора является одной из ключевых особенностей адаптивного подхода, ведь именно здесь происходит адаптация классификатора к особенностям данных из потока. При этом задействуется обратная связь по обучению, изображенная на рис. 1 и 3. Стоит добавить, что механизм, обеспечивающий классификацию, должен поддерживать добавление новых блоков в составной классификатор и замену существующих блоков.

Также необходим механизм, обеспечивающий однозначную идентификацию каждой точки данных из потока. Это позволит возвращать в общий поток точки данных, получившие метки классов только на этапе разметки, без потери данных.

**Заключение**

Инновационная технология, представленная в статье, устраняет основные недостатки современных методов классификации за счет адаптации к особенностям потока данных и введения нескольких петель обратной связи. Представленная технология, предполагающая получение и уточнение информации об особенностях распределения данных во время обработки, позволяет снизить зависимость эффективности методов машинного обучения от предварительного анализа данных и автоматически повышать качество классификации с течением времени, что можно рассматривать как конкурентное преимущество перед аналогами.

Описанный подход имеет следующие преимущества перед другими распространенными на данный момент методами классификации из области машинного обучения:

• обеспечивает работу с потоковыми данными, то есть позволяет производить обработку информационных объектов, которые не были отнесены ни к одному из существующих на тот момент классов, и не терять при этом полезную информацию;

• позволяет в полуавтоматическом режиме обновлять обучающую выборку, производя самостоятельно все этапы накопления и подготовки данных для дообучения и отдавая оператору только финальный этап разметки;

• осуществляет автоматический контроль качества обучающей выборки по заранее заданным параметрам;

• производит автоматическое обновление классификатора при появлении признаков устаревания в процессе поступления новых входных данных;

• реализует механизм борьбы с проблемой переобучения классификатора;

• поддерживает как полное, так и частичное обновление классификатора, приводящее к ускорению процесса обновления в целом и сохраняющее преемственность по обучающей выборке и непротиворечивость классификации до и после обновления.

Внедрение подобных технологий серьезно упростит обработку потоковых данных и станет важной ступенью к появлению адаптивных систем обработки данных, которые обучаются по мере поступления новой информации.

### Список литературы

1. Umesh Kokate, Arvind Deshpande, Parikshit Mahalle, Pramod Patil, Data Stream Clustering Techniques, Applications, and Models: Comparative Analysis and Discussion // Big Data and Cognitive Computing. 2018. Vol. 2. P. 32-62. DOI: 10.3390/bdcc2040032.

2. Частиков А.П., Урвачев П.М., Шевченко Д.В. Нейросетевой алгоритм распознавания паттернов в котировках фондовых бирж // Научный журнал КубГАУ. 2017. № 127(03). URL: http://ej.kubagro.ru/2017/03/pdf/20.pdf (дата обращения: 21.12.2023). DOI: 10.21515/1990-4665-127-020.

3. Ананченко И.В., Зудилова Т.В., Полин Я.А. О применимости алгоритмов кластеризации для борьбы со спамом в социальных сетях // Современные наукоемкие технологии. 2020. № 4-2. С. 190-194. DOI: 10.17513/snt.37995.

4. Демидова Л.А., Митин Г.В. Сравнительный анализ современных методов машинного обучения в контексте специфики их требований к обучающей выборке // Актуальные проблемы прикладной математики, информатики и механики: сборник трудов Международной научной конференции (г. Воронеж, 13-15 декабря 2021 г.). Воронеж: Научно-исследовательские публикации, 2022. С. 1547-1556.

5. Панов А.В., Митин Г.В. Сравнительный анализ современных методов машинного обучения в контексте специфики типов их выходных значений // Актуальные проблемы прикладной математики, информатики и механики: сборник трудов Международной научной конференции, (г. Воронеж, 12-14 декабря 2022 г.). Воронеж: Научно-исследовательские публикации, 2023. С. 1426-1435.

6. Турканов Г.И., Щепин Е.В. Классификатор Байеса для переменного количества признаков // Труды МФТИ. 2016. Т. 8, № 4(32). С. 8-12.

7. Парасич А.В., Парасич В.А., Парасич И.В. Формирование обучающей выборки в задачах машинного обучения // Информационно-управляющие системы. 2021. № 4(113). С. 61-70.