МЕЖДУНАРОДНЫЙ ЦЕНТР НАУЧНОГО СОТРУДНИЧЕСТВА
«НАУКА И ПРОСВЕЩЕНИЕ»

НАУКА и ПРОСВЕЩЕНИЕ
МЕЖДУНАРОДНЫЙ ЦЕНТР НАУЧНОГО СОТРУДНИЧЕСТВА

НАУЧНЫЕ ИССЛЕДОВАНИЯ 2026

СБОРНИК СТАТЕЙ XX МЕЖДУНАРОДНОЙ НАУЧНО-ПРАКТИЧЕСКОЙ КОНФЕРЕНЦИИ,
СОСТОЯВШЕЙСЯ 20 ФЕВРАЛЯ 2026 Г. В Г. ПЕНЗА

ПЕНЗА
МЦНС «НАУКА И ПРОСВЕЩЕНИЕ»
2026

ФИЗИКО-МАТЕМАТИЧЕСКИЕ НАУКИ

УДК 004.021

АЛГОРИТМ МАШИННОГО ОБУЧЕНИЯ СОПОСТАВЛЕНИЯ ДАННЫХ СИСТЕМЫ ФОРМИРОВАНИЯ ОПТИМАЛЬНЫХ ПРЕДЛОЖЕНИЙ НА ВЫБОР ИЗДЕЛИЙ ЭЛЕКТРОННОЙ КОМПОНЕНТНОЙ БАЗЫ В ПРОЦЕССАХ РАЗРАБОТКИ РАДИОЭЛЕКТРОННОЙ АППАРАТУРЫ

РУБЦОВ ЮРИЙ ВАСИЛЬЕВИЧ
генеральный директор
АО «ЦКБ «Дейтон»

Аннотация: В процессах разработки радиоэлектронной аппаратуры используются данные о параметрах и показателях изделий электронной компонентной базы. Такие данные в интегрированном виде можно получить из информационной справочной системы. Она выполняет сбор информации из различных источников, обрабатывает и предоставляет специалистам в виде оптимальной выборки данных.
Качество предоставляемых данных играет значимую роль и обеспечивается методами сопоставления на этапе нормализации. Для их выполнения используются детерминированные, вероятностные и с применением технологий искусственного интеллекта – машинного обучения, алгоритмы.
В настоящей статье показаны результаты исследования, разработки и использования алгоритма сопоставления данных методом машинного обучения для информационной справочной системы, разработанной АО «ЦКБ «Дейтон» и функционирующей для предприятий радиоэлектронной промышленности.
Ключевые слова: алгоритм, сопоставление данных, искусственный интеллект, машинное обучение, электронная компонентная база, радиоэлектронная аппаратура.

ALGORITHM FOR PROBABILISTIC COMPARISON OF DATA OF THE SYSTEM FOR FORMING OPTIMAL PROPOSALS FOR THE SELECTION OF ELECTRONIC COMPONENTS IN THE DEVELOPMENT PROCESSES OF RADIO ELECTRONIC EQUIPMENT

Rubtsov Yuri Vasilievich

Annotation: The development of electronic equipment utilizes data on the parameters and performance of electronic component products. This data can be obtained in an integrated form from an information reference system. It collects information from various sources, processes it, and provides it to specialists as an optimal data set.
The quality of the provided data plays a significant role and is ensured by matching methods at the normaliza-

tion stage. These methods utilize deterministic, probabilistic, and artificial intelligence (AI) algorithms, including machine learning.
This article presents the results of the research, development, and implementation of a machine learning data matching algorithm for an information reference system developed by JSC Central Design Bureau Dayton and operating for enterprises in the electronics industry.
Keywords: algorithm, data matching, artificial intelligence, machine learning, electronic component base, electronic equipment

Введение

Предприятия радиоэлектронной промышленности располагают всё большим количеством информации, из которой они получают необходимые данные, в том числе для разработки изделий электронной компонентной базы (ЭКБ), разработки радиоэлектронной аппаратуры (РЭА) и правильного применения ЭКБ. Время требует получать такую информацию с высоким уровнем актуальности и достоверности. Поэтому высококачественные данные имеют решающее значение. Такие данные предоставляют информационные справочные системы (Системы), в которых выполняется сбор информации, ее обобщение, анализ и формирование оптимальных предложений для применения. Особое место в Системе занимает нормализация данных, определение и описание представлены в [1-4]. Данные в Систему поступают из различных источников, в различных форматах, в разное время. Источниками информации в Систему являются: изготовители, поставщики и потребители ЭКБ; исследовательские, измерительные и испытательные организации; конструкторская и технологическая документация, протоколы результатов исследований, измерений, испытаний, применения, эксплуатации, и модернизации изделий ЭКБ.

Информация собирается в Системе, и для ее дальнейшей обработки необходима процедура сопоставления – отнесение наборов данных к конкретным изделиям ЭКБ.

В проведенных исследованиях определены три основные группы алгоритмов сопоставления данных:

1) точного сопоставления (детерминированные алгоритмы) полагаются на определенные шаблоны и правила [3];

2) неточное сопоставления (вероятностные алгоритмы) полагаются на методы для оценки вероятности отнесения наборов данных к одному и тому же изделию ЭКБ [2];

3) обучающие алгоритмы, образовались с развитием технологий искусственного интеллекта (ИИ). Применяется машинное обучение (МО). Используется широкий спектр признаков для расчета оценки сопоставления [4]. Выполняется определение весов сопоставления в соответствии с эталонным обучающим набором.

В настоящих исследованиях методы МО комбинируются с функциями строкового сходства в области сопоставления данных. Набор данных был предварительно обработан с помощью методов снижения размерности пар и набор функций сходства использовался для количественной оценки сходства между парами элементов наборов данных. Затем эти элементы использовались для обучения и валидации нейронной сети и дерева решений бустинга. Производительность была проанализирована по сравнению с другими доступными решениями для сопоставления. Дерево решений бустинга [5] и нейронная сеть, достаточно исследованы для апробирования и показали высокую производительность в применении задач сопоставления данных об изделиях ЭКБ.

Дерево решений бустинга (ДРБ) - это модель МО, которая объединяет несколько деревьев решений для повышения точности прогнозирования. Метод применяется путем последовательного построения деревьев, при этом каждое дерево исправляет ошибки своих предшественников. Этот подход использует сильные стороны отдельных деревьев решений для создания более надежной и точной модели сопоставления данных об изделиях ЭКБ.

Ценность данного исследования заключается в интеграции методов сравнения и решении на основе правил формирования данных об изделиях ЭКБ, которое реалистично как по объему, так и по соотношению ошибок к качественным данным.

Анализ исторических предпосылок настоящим исследованиям

С развитием информационной эпохи, за последние десятилетия, качественная обработка информации стала необходимостью для предприятий радиоэлектронной промышленности. Тем не менее, многие системы и процессы, которые интегрируют и хранят данные, подвержены ошибкам из-за человеческих факторов, ошибок в цифровизации и плохой интеграции.

Разработка изделий электронной техники — это область, которая особенно страдает от этих проблем, поскольку специалистам необходимо обрабатывать информацию, поступающую из различных источников, и добавлять ее в базы данных (БД). Результатом может быть дублирование наборов данных, вызванное тонкими и незначительными различиями в элементах, определяющих информацию об изделиях ЭКБ.

Область, занимающаяся этими проблемами, существует более пяти десятилетий и известна как сопоставление данных Многие доступные решения представляют собой модели на основе правил, разработанных специалистами в данной области [6,7].

В настоящих результатах исследований представлен подход, который использует специализированные знания в данной области, сочетает их с функциями сходства на основе расстояния для обобщения сравнения данных и выявления ошибок с помощью инструментов контролируемого обучения, технологий ICMH (Informatio Classification, Markingand Handling) – совокупности взаимосвязанных и взаимодействующих методов и инструментов, применяемых для решения задач сбора, обработки и анализа информации и получения достоверных данных в процессах сбора, обработки и анализа информации [8].

Эти инструменты показали высокую эффективность в нахождении скрытых паттернов [9] и корреляций в данных, которые можно использовать для их категоризации в различные группы. Более современные результаты исследований показали, что разработанные ранее инструменты оказались полезными для обнаружения ошибок данных в автоматизированных процессах управления данными об ИЭТ [10].

Данные и их предварительная обработка

Для выполнения исследований, описанных в этой статье, использовались данные об изделиях ЭКБ которые были предоставлены различными источниками информации: субъектами и объектами, идентифицирующие ее происхождение, доступными и дающими разрешение на ее использование Системой и обладающие достаточной определённостью. Информация достоверна, если она отражает истинное состояние изделия ЭКБ. Достоверную информацию Система получает от разработчиков и потребителей изделий ЭКБ, исследовательских, измерительных и испытательных организаций. Достоверная для Системы информация находится в конструкторской и технологической документации на изделия, в протоколах результатов исследований, измерений, испытаний, применения, эксплуатации, и модернизации изделий ЭКБ и РЭА.

Информация об изделиях ЭКБ состоит из наборов данных, которые могут иметь числовой и нечисловой вид. Под числовым видом понимаются строковые и числовые переменные, векторы, матрицы, многомерные массивы, константы. В нечисловом виде данные могут быть графиками, диаграммами, иллюстрациями, схемами. Обобщение и анализ числовых данных может проводиться с помощью математических операций и программного обеспечения (ПО), в то время как для нечисловых выполняются манипуляции с непосредственными сведениями об изделиях, а не их совокупностью в целом.

Чтобы получить очищенные обучающие данные, которые обеспечат лучшие результаты, пустые наборы данных об изделиях ЭКБ были удалены. Это допустимо, если процент удаления по отношению ко всем наборам данных остается низким (менее 2 процентов в данном исследовании).

Сопоставление данных требует, чтобы каждый набор данных сравнивался со всеми оставшимися наборами данных. Учитывая большое количество данных, это нецелесообразно с вычислительной точки зрения. Более того, такой подход к сопоставлению был бы очень неэффективным, поскольку некоторые наборы данных явно не соответствовали бы друг другу. Для решения этой проблемы используется блокировка. Она позволяет уменьшить размерность процесса сопоставления.

Дубликаты данных об изделиях ЭКБ, имеют по крайней мере один общий набор данных. Группируя наборы данных с общими значениями элементов, можно разделить их на непересекающиеся подмножества. Каждый набор данных затем сравнивается только с набором данных внутри этого подмножества. Это сокращает количество пар и значительно ускоряет процесс.

Сравнение наборов данных не является тривиальным из-за разнообразия типов элементов, которые их составляют. Элементы, связанные с датами, простыми числами, параметрами, показателями и текстовыми полями, имеют не только разные структуры и значения, но и их ошибки могут возникать из разных источников. Поэтому были определены несколько специализированных алгоритмов сопоставления для сравнения различных наборов данных несколькими способами.

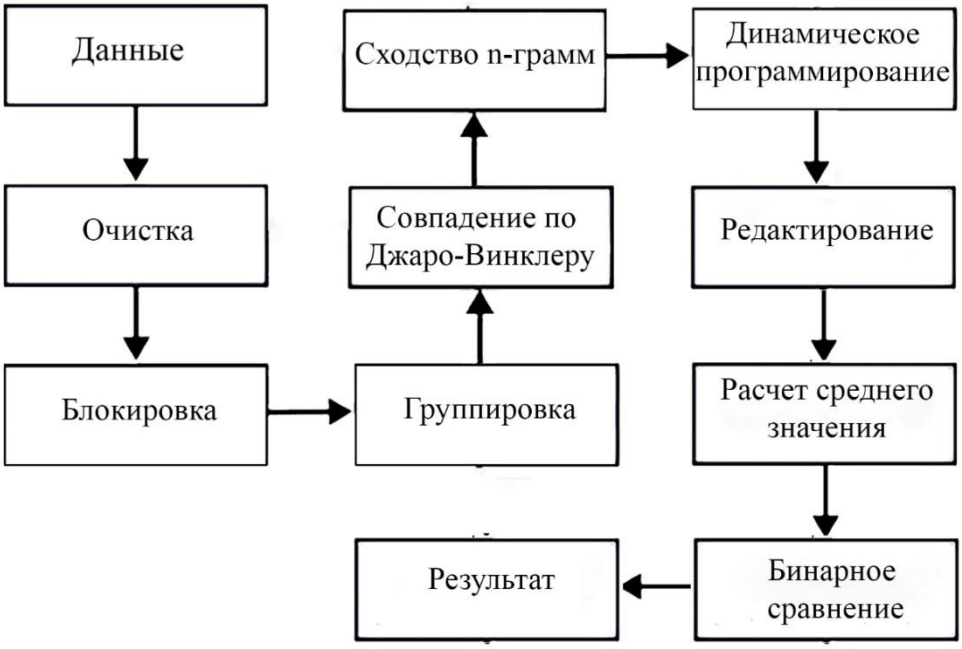

Рис.1. Алгоритм рабочего процесса предварительной обработки данных

Все алгоритмы сопоставления количественно оценивают степень сходства своим уникальным способом. Все выходные данные были нормализованы так, что единица указывает на точное совпадение, а ноль на отсутствие сопоставления. Рассматриваемые алгоритмы перечислены ниже:

1) Джаро-Винклера, в котором, строки, совпадающие вначале, получают более высокие оценки сходства;

2) сходства n-грамм Гжегожа Кондрака, сравнивает непрерывные последовательности из n символов, также известных как n-граммы. В данной работе использовались только 3- и 4-граммы;

3) Темпла Смита и Михаила Ватермана позволяет найти соответствующие последовательности символов с помощью динамического программирования, подхода, который раскладывает задачи на более простые подзадачи;

4) Владимира Левенштейна, определяет количество операций с одиночными символами (вставка, удаление или замена), необходимых для преобразования одной строки в другую. С учетом расширения Фредерика Дамерау, при котором перестановка двух смежных символов также рассматривалась как операция редактирования;

5) Гари Бенсона, для нахождения длиннейшей общей подстроки, которая является подстрокой обеих сравниваемых строк;

6) Алваро Монжа и Чарлеса Элкана, для применения к строке, которая состоит из нескольких строк, разделенных пробелами;

7) бинарное сравнение, для сравнивания двух значений элементов наборов данных и получение единицы, если они полностью совпадают, и ноль, если хотя бы один символ отличается.

Вышеупомянутые алгоритмы были выбраны таким образом, чтобы охватить как можно больше различных типов наборов данных. Результатом является вектор значений сопоставления для каждой пары набора данных об ЭКБ, который будет использоваться в качестве входных данных для алгоритмов МО. Все пары наборов данных затем маркируются с учетом результатов сопоставления. Алгоритм рабочего процесса предварительной обработки данных представлен на рис. 1.

Подходы машинного обучения

Сопоставление наборов данных является задачей классификации, и размеченные пары элементов позволяют использовать методы контролируемого обучения. В этом контексте были выбраны следующие два метода:

1) ДРБ объединяет несколько обучающих моделей - классических решающих деревьев (КРД), и комбинирует их в одну сильную модель. КРД обучаются последовательно на выборке, веса которой рассчитываются в зависимости от точности предыдущего КРД. ДРБ показывает точность и меньшую подверженность переобучению. В данном исследовании использовалась улучшенная форма бустинга, - градиентный бустинг.

В наших исследованиях каждый уровень обучения не только обучается на текущих данных, но также на остатках предыдущего обучения, разнице между предсказанным и целевым значением. Это добавление приводит к более стабильной и быстрой сходимости обучения.

Три тысячи наборов данных прошли обучение с использованием модифицированного метода наименьших квадратов [11], при коэффициенте скорости обучения 0.1 и максимальной глубине узла 4 (путь от корневого узла дерева). Количество уровней обучение, коэффициент скорости обучения и максимальная глубина узла были определены с помощью перекрестной валидацией через поиск гиперпараметров.

Перекрестная валидация — это метод оценки ДРБ и ее поведения на подготовленных данных. При оценке ДРБ данные разбивались на k частей. Затем на k−1 частях данных производилось обучение ДРБ, а оставшаяся часть данных использовалась для тестирования. Процедура повторялась k раз; в итоге каждая из k частей данных использовалась для тестирования. Процессы обучения модели ДРБ представлены на рис.2.

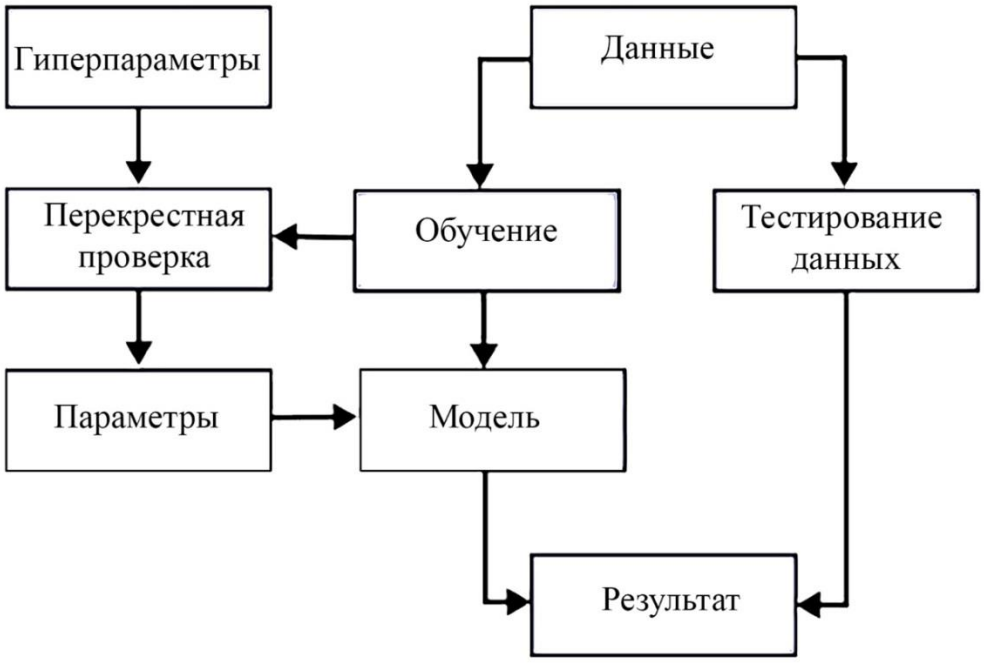

Рис.2. Среднее значение ROC – кривой по всем выборкам 5-кратной кросс-валидации с применением ДРБ

Гиперпараметры имели в настоящих исследованиях решающее значение для достижения высокой производительности и точности ДРБ. Они устанавливались до начала процесса МО и определяли как саму структуру ДРБ, так и способ МО. Их правильный выбор значительно повлиял на результаты обучения, потому что неправильно подобранные значения гиперпараметров могут привести к нежелательным результатам.

Способность ДРБ к сопоставлению данных — ключевой аспект, определяющий ценность ДРБ в решении задач Системы. Гиперпараметры влияют на эту способность, играя роль регулятора ДРБ. Правильно подобранные гиперпараметры помогают сбалансировать между переобучением и недообучением ДРБ.

Нейронная сеть (НС) — это алгоритм МО, построенный на основе человеческого мышления. Она основана на коллекции связанных узлов, подобно человеческим нейронам. Подобным образом значения признаков передаются узлам входного слоя, которые, в свою очередь, передают их узлам следующего скрытого слоя и так далее. Входной слой в наших исследованиях состоит из 20 узлов, что соответствует количеству рассматриваемых входных признаков. Сеть имеет два скрытых слоя, каждый из которых содержит 30 узлов. Все слои используют активационную функцию. Выходной слой состоит из одного узла, который использует сигмоидальную активацию для вычисления вероятности дублирования. Обучение проводится с использованием алгоритма оптимизации и функции потерь бинарной кросс-энтропии.

Для обеспечения оптимальной производительности обе модели были подвергнуты настройке гиперпараметров с использованием 5-кратной кросс-валидации на наборе данных.

Результаты исследований

В этом разделе показано, как модели обучены и проверены на созданном наборе данных. Представлены две различные стратегии для настройки моделей и оценки их ошибок на имеющемся наборе данных.

4.1. Проверенной стратегией обучения и валидации моделей является пятикратная перекрестная валидация.

Набор данных уменьшался в недублируемых парах элементов набора данных до соотношения один к одному с парами дублирующих элементов набора данных. Затем данные разбивались на пять равных по размеру подмножеств, из которых четыре используются для обучения, а одно для валидации. Модель обучалась и проверялась пять раз, так что каждое подмножество использовалось один раз для валидации. В проведенных исследованиях использованы пространства сопоставления данных: Пл — истинные положительные, Лп — ложные положительные, Ло — ложные отрицательные, От — истинные отрицательные.

Для оценки результатов использована ROC-кривая (Receiver Operating Characteristic curve). Площадь под ROC является достаточной метрикой производительности для сравнения между моделями в проведенных исследованиях [12]. ROC-кривая иллюстрирует производительность модели при всех возможных порогах классификации.

Для оценки результатов, ROC-кривая, представляется на графике. На оси X такого графика указаны значения показателя ложно положительных результатов (ЛПР). Он определяет долю ложно положительных результатов относительно всех отрицательных результатов и вычисляется как (1):

$$\text{ЛПР} = \frac{\text{Лп}}{\text{Лп}+\text{От}} \quad (1)$$

ЛПР оценивает способность модели правильно определять отрицательные результаты сопоставления.

На оси Y данного графика указаны значения показателя истинные положительные результатов (ИПР). Он определяет долю истинных положительных результатов относительно всех положительных результатов и вычисляется как (2):

$$\text{ИПР} = \frac{\text{Пл}}{\text{Пл}+\text{Ло}} \quad (2)$$

ИПР определена как чувствительность. Она оценивает способность модели правильно определять истинные положительные результаты сопоставления.

Кривая ROC строится с ИПР по оси Y и ЛПР по оси X для различных пороговых значений. Чем ближе кривая ROC находится к верхнему левому углу графика, тем лучше производительность модели. Идеальная ROC проходит через верхний левый угол (ИПР = 1, ЛПР = 0), что означает высокую чувствительность и низкую частоту ложных срабатываний.

В настоящих исследованиях применена мера, которая позволяет суммировать производительность модели одним числом, измеряя площадь под кривой ROC — AUC (Area Under the ROC Curve). AUC колеблется от 0 до 1. Более высокое значение AUC указывает на более высокую производительность модели. AUC равный 0.5 указывает на отсутствие дискриминационной способности модели, тогда как AUC равный единице означает на идеальную производительность модели.

AUC показывает инвариантность к порогу классификации и масштабу предсказуемости. AUC не зависит от масштаба вероятностей, которые генерирует модель. Например, две модели могут выдавать предсказания в различных масштабах, одна — в виде вероятностей от 0 до 1, а другая — в виде более широкого диапазона значений. Несмотря на эти различия, AUC как мера будет одинаковым, если порядок ранжирования случаев от наиболее вероятного положительного до наиболее вероятного отрицательного сохраняется.

Для построений графика ROC и расчета AUC использованы средства программирования Python:

1) подготавливаются данные и обучается модель. Используется логистическая регрессию для бинарной классификации;

2) импортируются необходимые библиотеки и загружаются данные;

3) обучается модель и делается предсказания вероятностей;

4) для расчета значений ИПР и ЛПР используем функция roc_curve, а затем строится ROC – кривая;

5) для оценки качества модели рассчитывается площадь под ROC-кривой (AUC), используется функция roc_auc_score.

На рис. 3 показано среднее значение ROC – кривой по всем выборкам 5-кратной кросс-валидации с применением ДРБ.

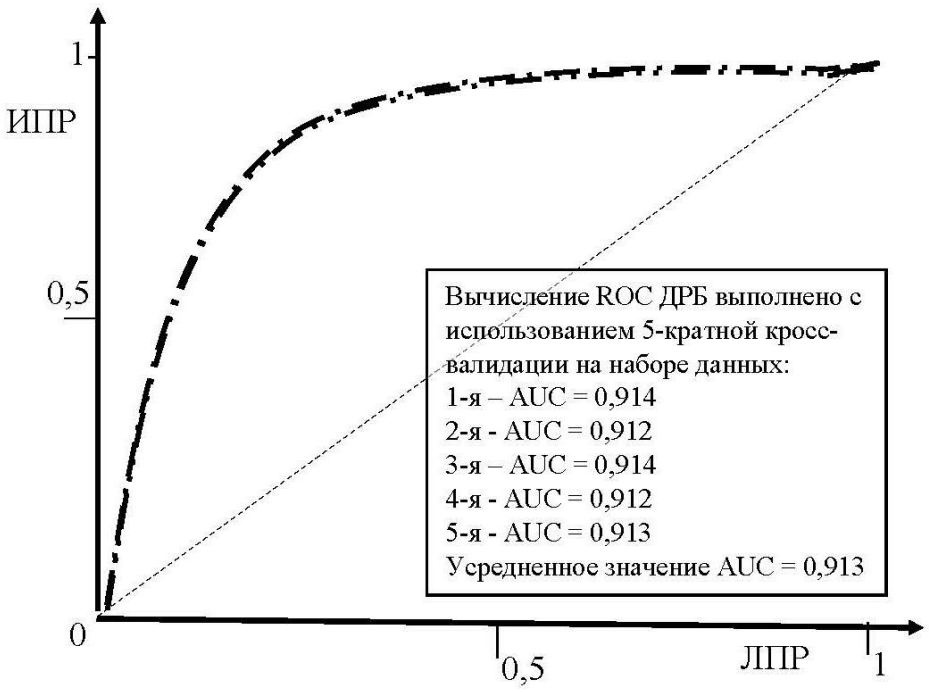

Рис.3. Среднее значение ROC – кривой по всем выборкам 5-кратной кросс-валидации с применением ДРБ

На рис. 4 показано среднее значение ROC – кривой по всем выборкам 5-кратной кросс-валидации с применением НС.

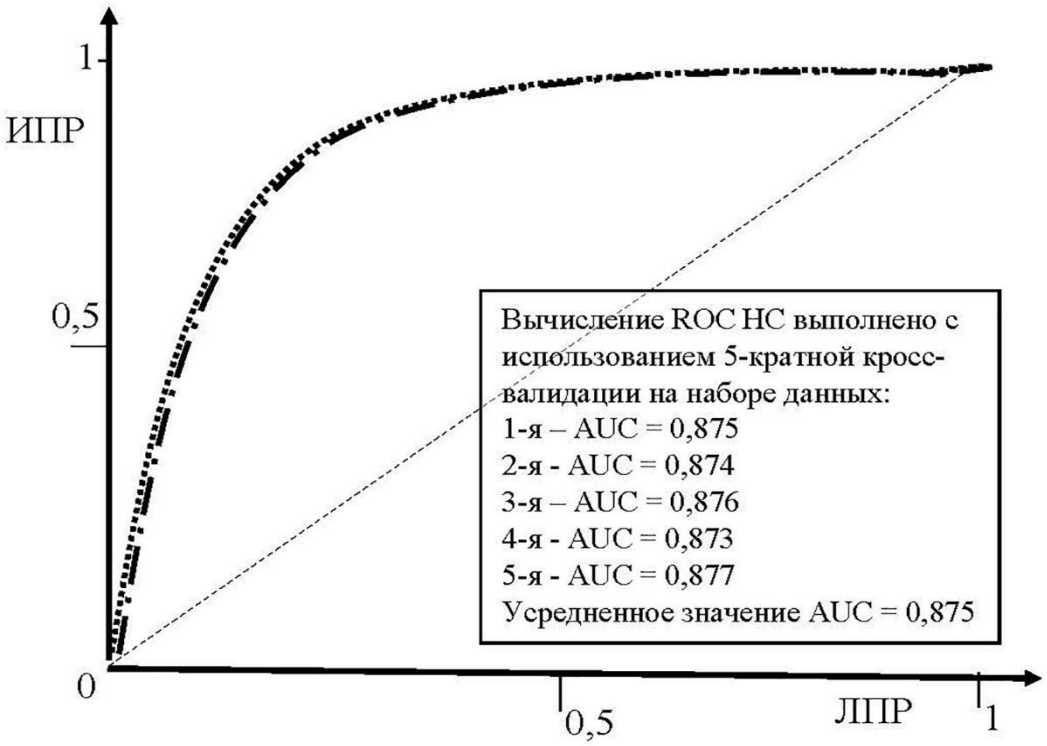

Рис.4. Среднее значение ROC – кривой по всем выборкам 5-кратной кросс-валидации с применением НС

Результаты анализа изменения на графике ROC – кривой и значений AUC на рис.3,4 показывают, что ДРБ имеет лучшие результаты чем НС плаченые за счет компромисса между чувствительностью (1) и уровнем ложных срабатываний (2).

Диагональная линия, проходящая через график, представляет собой ограничитель, который должен быть ниже ROC. Диагональная линия, предсказывает положительные и отрицательные результаты с равной величиной. Это наихудший сценарий сопоставления данных.

4.2. Нехватка данных для валидации и несбалансированная валидация

Далее, разработанному алгоритму были предоставлены обучающие данные, которые все еще недообработаны, но проверяются на данных с исходным классом несбалансированности. Набор неполных данных, потупивший от одного источника используется для валидации в его несбалансированной форме. Данные, соответствующие другим источникам, используются как обучающие. Процедура повторяется до тех пор, пока данные от каждого источника не будет использованы один раз для валидации. Результаты по всем источникам накапливаются и оценивался как один набор тестовых данных. Исследования показали, что ROC-кривые на графике могут вводить в заблуждение в случае бинарной классификации с несбалансированными наборами данных, так как преобладающие истинные негативные результаты могут снижать уровень ложных положительных результатов, в то время как точность (3) может оставаться низкой.

$$\text{Т} = \frac{\text{Пл}}{\text{Пл}+\text{Лп}} \quad (3)$$

Низкая точность приведет к тому, что пары элементов в наборах данных, в большинстве будут состоять из ложных положительных результатов. Поэтому AUC добавляется в качестве метрики производительности, как рекомендуется для несбалансированных наборов данных [13,14].

Более того, важно учесть, что как Лп, так и Ло влияют на результат различными способами; первые требуют дополнительных процедур обработки данных, в то время как вторые допускают пропущенные данные, что соответствует потерянной информации. Относительная важность этих двух событий меняется в зависимости от источников данных. Чтобы учесть эту разницу в важности введена F-мера, как среднее гармоническое взвешенное (4).

$$F = (1 - \beta)^2 \frac{\text{ИПР} \cdot T}{\beta^2 \cdot (\text{ИПР} + T)} \quad (4)$$

F-мера изменяется таким образом, что ИПР важнее Т в β количество раз. Таким образом, если ложноотрицательные результаты в 5 раз важнее, чем ложноположительные, то модель с наилучшей F-мерой та, в которой β = 5.

На рис. 5 представлено изображение зависимости значений F-меры от порогов вероятности для β = 5. Для ДРБ максимальное значение равно 0,6, для НС 0,3, что показывает значительную эффективность ДРБ по сравнению с НС.

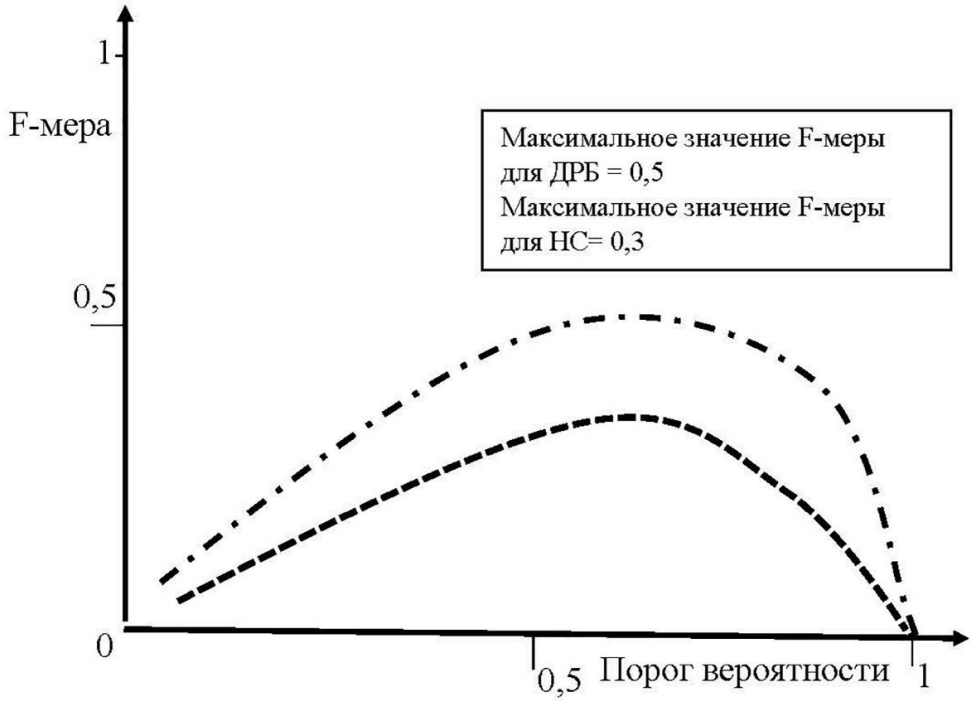

Рис.5. Значение F-меры для β = 5 и различных порогов вероятности сопоставления

Для β = 1, на рис. 6 представлено изображение зависимости значений F-меры от порогов вероятности. Максимальное значение уменьшилось в 2 раза по сравнению с β = 5. Но при этом остается большее значение для ДРБ, по сравнению с НС.

F-мера достигает максимума при ИПР и Т, равными единице, и близка к нулю, если один из аргументов близок к нулю.

В настоящих исследованиях применена логистическая функция потерь для оценки разности между прогнозируемыми и истинными значениями, для каждого сопоставляемого набора данных (5).

$$\text{ФП} = -\frac{1}{n} \cdot \sum_{i=1}^{i} (y_i - \log(\hat{y}_i) + (1 - y_i) - \log(1 - \hat{y}_i)) \quad (5)$$

где: n – кратность кросс-валидации,

$\hat{y}_i$ — прогнозируемый результат на $_i$-том элементе набора данных,

y — Пл для i-го элементе набора данных, а i количество элементов набора данных.

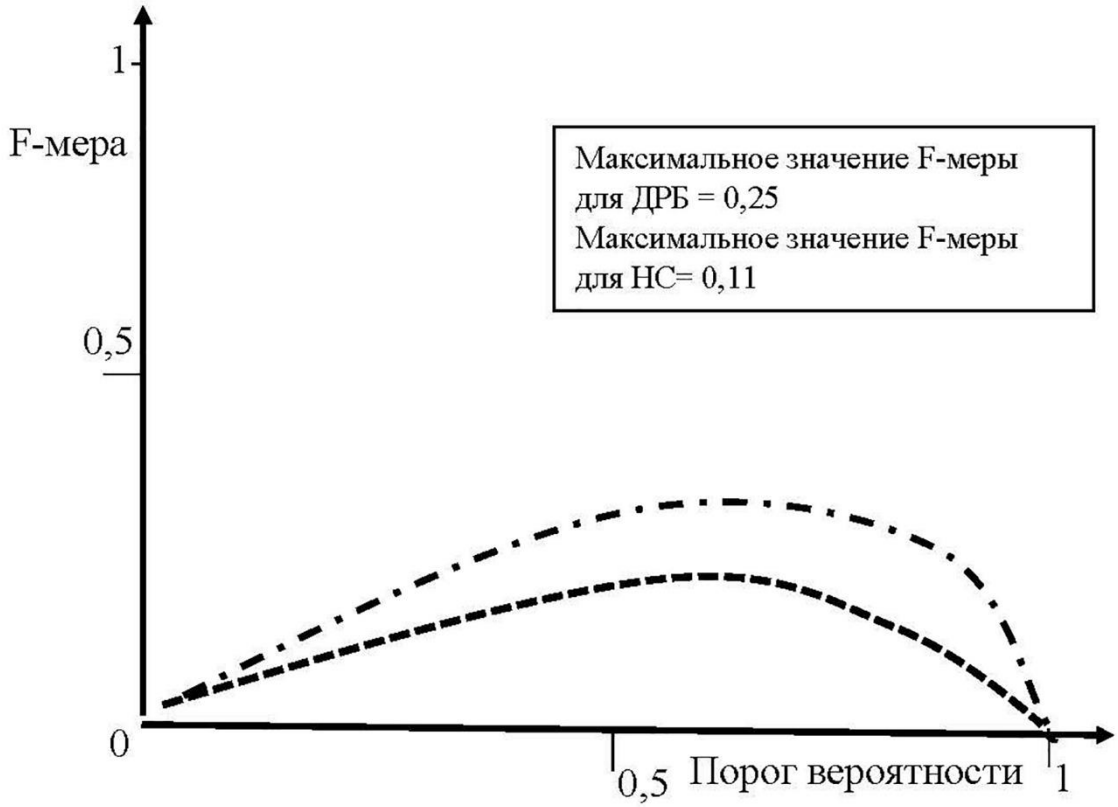

Рис.6. Значение F-меры для β = 1 и различных порогов вероятности сопоставления

Минимизация ФП представляется как задача максимизации Т путем штрафа за ошибки при проведении МО.

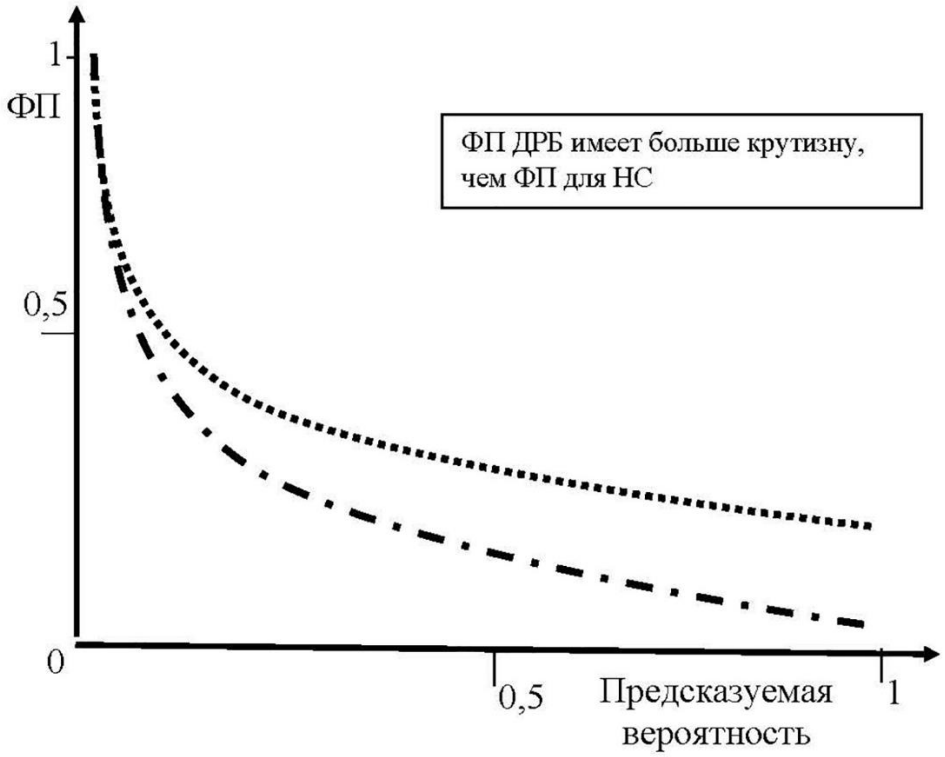

Рис.7. Диапазон возможных значений логистической функции потерь для ДРБ и НС

На рис.7 показан диапазон возможных значений логистической функции потерь для ДРБ и НС. Когда прогнозируемая вероятность приближается к 1, логистическая функция потерь медленно уменьшается. При уменьшении прогнозируемой вероятности она быстро возрастает. ФП ДРБ более оптимальна, имеет большую крутизну чем для НС.

Заключение

В данной работе мы показали, что результаты сопоставление данных об изделиях ЭКБ могут быть значительно улучшены с использованием методов контролируемого обучения. Были взяты за основу фундаментальные методы ДРБ и НС. Поскольку источники данных об изделиях ЭКБ разные, фактически функционируют с разными форматами, невозможно количественно оценить эффективность любого инструмента с помощью одного значения. В целом, ДРБ показало наилучшие результаты и может обеспечить значительное улучшение результатов сопоставления данных. НС имеет производительность ниже, но обеспечивает производительность для более широкого диапазона уровней вероятностей, что делает её более гибким инструментом.

Представленная работа демонстрирует перспективы для будущих исследований. Функции сходства, использованные для построения признаков, в основном основаны на текстовых и числовых данных. Добавление признаков, которые количественно оценивают сходство между данными другого типа (графики, схемы, условные графические обозначения), могло бы увеличить дискриминационную способность моделей. Другим дополнением было бы исследование возможностей неконтролируемого обучения.

Дисбаланс результатов был решен путем уменьшения обучающих данных, что остается проблемой, приводящей к большому числу ложноположительных результатов. Исследования более продвинутых методов могут еще больше повысить эффективность инструментов на основе МО.

При использовании ROC-кривой и AUC иногда возникают ошибки, которые могут привести к неправильной интерпретации производительности модели. Одной из основных ошибок при использовании AUC является игнорирование дисбаланса результатов. Когда один результат значительно превышает другой, модель может быть смещена в сторону большего результата.

Проведены исследования и определены способы минимизации ошибок при применении ROC-кривой и AUC:

1) использование методов балансировки обучающей выборкой, таких как взвешивание результатов в процессах обучения, использование техник корректировки обучающей выборки с целью балансировки распределения данных в исходной выборке;

2) пересчет метрик в дополнение к AUC, таких как точность, чувствительность и F-мера, которые могут дать более полную картину при несбалансированной обучающей выборке.

AUC часто интерпретируется как вероятность того, что случайная положительная выборка будет оценена моделью выше, чем случайно отрицательная выборка. Ошибка заключается в том, что высокий AUC не всегда указывает на высокую эффективность модели во всех ситуациях, особенно если классы сильно не сбалансированы. Для устранения используется комбинированные подходы при оценке модели, включая анализ ROC-кривой для определения оптимального порога, который обеспечивает приемлемый баланс между ИПР и ЛПР.

Форма ROC-кривой зависит от выбора порогового значения, это может привести к ошибочным выводам о производительности модели, особенно если порог выбран не оптимально. Для минимизации ошибок необходимо выполнить:

1) изучение ROC-кривой для разных порогов, чтобы понять, как изменение порога влияет на ИПР и ЛПР;

2) выбор порога, который обеспечивает наилучший компромисс между обнаружением положительных и отрицательных случаев в соответствии с задачами сопоставления.

Список источников

1. Рубцов Ю.В. Алгоритм сопоставления для нормализации данных системы формирования оптимальных предложений на выбор изделий электронной компонентной базы в процессах разработки радиоэлектронной аппаратуры // Автоматизация в промышленности. Алгоритмическое и программное обеспечение. №7 2025 - URL: https://www.deyton.ru/doc/rub09.07.2025.pdf (дата обращения: 10.02.2026).

2. Рубцов Ю.В. Алгоритмы вероятностного сопоставления данных системы формирования оптимальных предложений на выбор изделий электронной компонентной базы в процессах разработки радиоэлектронной аппаратуры. Актуальные научные исследования, сборник статей XXXII Международной научно-практической конференции. 15 февраля 2026 - URL: https://www.deyton.ru/doc/stat19.02.2026.pdf (дата обращения: 10.02.2026).

3. Рубцов Ю.В. Алгоритмы точного сопоставления данных системы формирования оптимальных предложений на выбор изделий электронной компонентной базы в процессах разработки радиоэлектронной аппаратуры. Научное обозрение: актуальные вопросы теории и практики, сборник статей XXI Международной научно-практической конференции. 10 февраля 2026 года)- URL: https://www.deyton.ru/doc/stat12.02.2026.pdf (дата обращения: 10.02.2026).

4. Рубцов Ю.В. Оценка метода машинного обучения для системы автоматизированного выбора компонентной базы радиоэлектронной аппаратуры. Автоматизация и измерения в машиноприборостроении: научный журнал, №3, 2025 - URL: https://www.deyton.ru/doc/stat25.06.2025.pdf (дата обращения: 10.02.2026).

5. Chang Yu, Fang Liu, Jie Zhu, Shaobo Guo, Yifan Gao, Zhongheng Yang, Meiwei Liu, Wuhan,Hubei, Qianwen Xing. Gradient Boosting Decision Tree with LSTM for Investment Prediction. 5th Asia-Pacific Conference on Communications Technology and Computer Science, 2025 - URL: https://www.computer.org/csdl/proceedings-article/acctcs/2025/246300a057/29QhWXzkr1C (дата обращения: 10.02.2026).

6. Ravi Gupta. The Future of Data Matching: AI, Machine Learning, and Beyond. Digital Transformation 2025 - URL: https://ve3.global/blog/the-future-of-data-matching-ai-machine-learning-and-beyond (дата обращения: 10.02.2026).

7. Tristan Thommen. AI Data Matching: Unify Your Information for Better Decisions. Koncile SAS 2025 - URL: https://www.koncile.ai/en/ressources/data-matching-unify-your-data-for-smarter-decisions (дата обращения: 10.02.2026).

8. Дормидошина Д.А., Савин М.Л., Рубцов Ю.В., Применение ICMH в процессах сбора, обработки и анализа информации о надежности изделий микроэлектроники. «Нано- и микросистемная техника», Том 22, №9, г. Москва 2020 - URL: https://www.deyton.ru/doc/ISSN18138586.pdf (дата обращения: 10.02.2026).

9. Зуенко А.А., Фридман О.В., Обзор методов поиска частых паттернов для интеллектуального анализа данных. Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра РАН. Труды Кольского научного центра РАН. Серия: технические науки. Том: 15 №: 3 2024 - URL: http://kolanord.ru/html_public/periodika/Trudy-KNC/2024/Trudy_KNC_Vyp%2015_2024_N3_Tehnicheskie-nauki/2/ (дата обращения: 10.02.2026).

10. Свидетельство о государственной регистрации программы для ЭВМ №2022668891 Российская Федерация, Программа управления данными об изделиях электронной техники: заявка №2022667557, дата поступления 27.09.2022: дата государственной регистрации в Реестре программ для ЭВМ 13.10.2022/ Рубцов Ю.В., Дормидошина Д.А., Криницкий В.В., Владимиров А.И., Окунев К.Е.; заявитель АО «ЦКБ «Дейтон».

11. Матвеев М.Г., Сирота Е.А., Исследование решения задачи параметрической идентификации моделей распределенных динамических процессов. Актуальные проблемы прикладной математики, информатики и механики. Сборник трудов Международной научной конференции. Воронеж, 2021 г - URL: https://www.amm.vsu.ru (дата обращения: 10.02.2026).

12. А.А. Софронов, А.А. Рябов, А.А. Горбунов, Л.Ю. Ротков. Обнаружение сетевых аномалий в

реальном трафике с использованием метода машинного обучения Random Forest. Труды XXIX Научной конференции по радиофизике, ННГУ, 2025 - URL: https://rf.unn.ru/wp-content/uploads/sites/21/2025/10/rf-conf-2025-information-systems.pdf (дата обращения: 10.02.2026).

13. Частикова В.А., Лях А.Р. Методы машинного обучения в задачах балансировки данных. Кубанский государственный технологический университет. Электронный сетевой политематический журнал "Научные труды КубГТУ". №: 4 2021 - URL: https://ntk.kubstu.ru/data/mc/0083/4087.pdf (дата обращения: 10.02.2026).

14. Gaurav Sharma, Thomas Mathew. One-Sided and Two-Sided Tolerance Intervals in General Mixed and Random Effects Models Using Small-Sample Asymptotics. Journal of the American Statistical Association 2025 - URL: https://www.researchgate.net/publication/241722180_One-Sided_and_Two-Sided_Tolerance_Intervals_in_General_Mixed_and_Random_Effects_Models_Using_Small-Sample_Asymptotics (дата обращения: 10.02.2026).

НАУЧНОЕ ИЗДАНИЕ

НАУЧНЫЕ ИССЛЕДОВАНИЯ 2026

Сборник статей
Международной научно-практической конференции
г. Пенза, 20 февраля 2026 г.
Под общей редакцией
кандидата экономических наук Г.Ю. Гуляева

Подписано в печать 21.02.2026.
Формат 60×84 1/16. Усл. печ. л. 11,5

МЦНС «Наука и Просвещение»
440062, г. Пенза, Проспект Строителей д. 88, оф. 10
www.naukaip.ru

16+