

УДК 007.5:519.816:681.3.016

doi: 10.15622/rcai.2025.080

## ИСПОЛЬЗОВАНИЕ МЕТОДОВ АНСАМБЛИРОВАНИЯ ДЛЯ ПОВЫШЕНИЯ ТОЧНОСТИ КЛАССИФИКАЦИИ ОБЪЕКТОВ, СОДЕРЖАЩИХ ШУМ<sup>1</sup>

М.В. Фомина (*m\_fomina2000@mail.ru*)

Г.В. Швец (*ShvetsGV@mpei.ru*)

Национальный исследовательский университет «МЭИ», Москва

Рассматривается проблема интеллектуального анализа зашумленных данных, полученных из реальных источников. Предлагается метод организации системы классификации объектов, подверженных влиянию шума различного типа. В основу положено сочетание ансамблевых методов с нейросетевым подходом. Предложена архитектура нейросети, решающей поставленную задачу. Приводятся результаты машинного эксперимента, даётся оценка влиянию шумов различного типа на результаты классификации.

**Ключевые слова:** классификация данных, модели шума, нейронная сеть, ансамблевый метод, предобработка данных.

### Введение

Системы искусственного интеллекта, способные оказать помощь людям в их профессиональной сфере при принятии ответственных решений, требуют обработки и анализа больших потоков данных. В настоящее время вместе с увеличением объема информации растет также и проблема её качества. Реальные данные, получаемые из различных областей, неизбежно содержат шум: это искажения данных, или пропуски в данных [Goodfellow et al., 2016], [Гафаров и др., 2018].

Присутствие различного рода шума в данных оказывает существенное влияние на результаты их анализа, особенно в задачах классификации данных, где требуется высокая точность отнесения рассматриваемых объектов к различным категориям (классам). Эта проблема становится осо-

---

<sup>1</sup> Работа выполнена при финансовой поддержке РНФ (проект № 24-11-00285) <https://rscf.ru/project/24-11-00285>.

бенно важной в таких областях, где ошибка классификации может иметь крайне серьезные последствия, например, при разработке систем поддержки принятия решений [Еремеев и др., 2021].

В последние годы нейронные сети хорошо зарекомендовали себя как мощный инструмент решения разнообразных задач машинного обучения, включая классификацию данных. Благодаря способности изучать сложные нелинейные зависимости и высокой обобщающей способности, нейронные сети хорошо использовать для задач анализа зашумленных данных [Москвитин и др., 2023]. Современные архитектуры нейронных сетей способны демонстрировать устойчивость к определенным видам шума, выделяя значимые паттерны и игнорируя случайные искажения.

В работе рассматривается проблема улучшения классификации зашумленных данных при высоких уровнях шума за счёт использования комплексного подхода – интеграции ансамблевых методов и использования нейросети определённого типа. Приведены результаты, позволяющие оценить влияние шума различных типов на успешность решения задачи классификации. Рассмотрена организация программного комплекса, реализующего предложенные методы и алгоритмы.

## **1. Проблема классификации объектов в интеллектуальных системах**

Задача классификации данных различного типа, является одной из фундаментальных задач машинного обучения, целью которой является отнесение объектов к одному из определенных классов на основе анализа признаков, отражающих свойства объектов [Вагин и др., 2008].

Для успешной классификации объектов требуется прежде всего решить задачу построения набора классификационных правил – классификатора, который, по сути, является алгоритмом, дающим ответ на вопрос, принадлежит ли очередной пример классу, или не принадлежит. Построение классификатора, или классификационной модели, основано на анализе конкретных примеров с целью поиска свойств, наиболее характерных для них. Примеры, которые используются для анализа, представляют собой объекты, которые могут быть описаны набором свойств (атрибутов); такое описание называется признаковым описанием.

Известны различные подходы и модели для построения классификаторов; это продукционные правила, деревья решений, методы, основанные на теории приближённых множеств, нейронные сети. Важно, что ни один из методов не является универсальным и обычно показывает наилучшие результаты на наборах данных определённого типа [Еремеев и др., 2023].

Построение хорошей классификационной модели выполняется на примерах, которые образуют обучающую выборку. В перечисленных выше методах объекты обучающей выборки снабжаются меткой – именем класса.

Если обучающая выборка размечена, речь идёт об обучении «с учителем», классификационные модели, работающие с неразмеченными данными, решают задачу обучения «без учителя».

Следует отметить, что успешность, или полезность построенного классификатора определяется его способностью правильно распознавать новые, не вошедшие в обучающую выборку примеры. Также полезным свойством классификационной модели будет устойчивость к шуму в данных. Для оценки этого свойства используется оценка *Точность классификации* или *Accuracy*, основной показатель оценки производительности модели, который измеряет долю верных ответов среди всех сделанных предсказаний [Goodfellow et al., 2016].

Для оценки эффективности классификационной модели используются следующие наборы данных:

- Обучающая выборка (Training set) – выборка, используемая для обучения модели.
- Тестовая выборка (Test set) – выборка, используемая для окончательной оценки производительности модели.
- Валидационная выборка (Validation set) – выборка, используемая для настройки гиперпараметров и выбора лучшей модели.

Шум оказывает комплексное негативное воздействие на все аспекты процесса машинного обучения, начиная от качества данных, заканчивая практическим применением моделей. Разработка методов, устойчивых к различным типам шума является важной задачей для создания надежных систем классификации.

## 2. Модели шума и их характеристики

Как уже говорилось ранее, шум – это нежелательные искажения данных, которые влекут за собой снижение успешности работы алгоритмов классификации [Gonsales et al., 2018]. Шум в данных представляет собой непредсказуемые вариации, поскольку искажению подвергаются истинные значения. В контексте задач классификации можно выделить несколько основных типов шума: гауссовский, импульсный, равномерный, пропущенные значения, мультипликативный, шум типа «соль и перец». Шум каждого типа по-своему влияет на результаты классификации и требует специализированных подходов для эффективной обработки данных [Fomina et al., 2013].

В рамках исследования был разработан программный комплекс, одной из функций которого является искусственное внесение в данные шума различного типа. Была реализована поддержка следующих ключевых типов шума.

*Гауссовский шум* (Gaussian noise). Является наиболее распространенным типом шума в реальных данных и характеризуется нормальным распределением значений с нулевым математическим ожиданием.

*Импульсный шум* (Impulse noise). Характеризуется редкими, но значительными по амплитуде выбросами, которые изменяют истинные значения. Импульсный шум особенно опасен тем, что может спровоцировать радикальное искажение ключевых характеристик отдельных объектов, потенциально переводя их в область другого класса.

*Равномерный шум* (Uniform noise). Проявляется в случайных добавках к данным, которые имеют равномерное распределение в заданном интервале [Gonsales et al., 2018].

*Пропущенные значения* (Missing values). Данный тип представляет собой особый шум, при котором некоторые значения признаков по какой-либо причине отсутствуют. При этом может быть известна  $p$  – вероятность того, что некоторое значение отсутствует.

Пропущенные значения представляют особую сложность для алгоритмов машинного обучения, не способных работать с неполными данными [10]. Это требует дополнительной обработки с использованием специальных методов восстановления отсутствующего значения, что вносит дополнительную неопределенность в процесс анализа.

*Шум типа «Соль и перец»* (Salt and pepper noise). Этот тип шума представляет собой случайную замену значений на экстремальные минимумы (*перец*) или максимумы (*соль*)  $x_{\min}$ ,  $x_{\max}$ . Данный шум является частным случаем импульсного шума, при этом величины  $x_{\min}$ ,  $x_{\max}$  – это минимальное и максимальное значения признака. Также известны вероятности проявления «перца» и «соли» соответственно [Gonsales et al., 2018]. Такой шум создает резкие аномалии в данных, что может привести к значительным искажениям признаковового пространства. В задачах классификации подобный шум может создавать ложные кластеры или искажать истинные границы между классами.

### **3. Расширение возможностей нейронных сетей с помощью ансамблевых методов**

Нейронные сети представляют собой мощный инструмент для создания моделей, способных эффективно аппроксимировать сложные нелинейные зависимости между входными признаками и целевыми переменными [Гафаров и др., 2018].

За счет применения специальных методов и техник, которые направлены на улучшение обобщающей способности и снижение чувствительности к искажениям в данных можно значительно улучшить эффективность работы нейросетей. Следующие ключевые подходы позволяют сде-

лать нейронные сети более устойчивыми к шуму: регуляризация, нормализация, использование глубоких нейронных сетей с резидуальными соединениями [He et al., 2016].

Регуляризация представляет собой набор методов, которые направлены на предотвращение переобучения модели путем наложения дополнительных ограничений на гиперпараметры или структуру нейронной сети [Goodfellow et al., 2016]. При наличии зашумленных данных регуляризация приобретает особую важность, поскольку предотвращает запоминание зашумленных данных нейросетью.

Методы нормализации активаций в нейросетях стабилизируют процесс их обучения, ускоряют сходимость и повышают устойчивость к различным типам шума.

В качестве базовой модели нейронной сети в работе предлагается использовать сеть с резидуальными соединениями, где передаточная функция  $F(x)$  представляет последовательность двух взвешенных слоёв. При наличии зашумленных данных использование такой структуры нейросети с резидуальными соединениями даёт ряд преимуществ. Перечислим их.

1. Улучшенное распространение градиентов. При использовании резидуальных соединений локальный градиент функции потерь относительно входа слоя гарантированно не затухает даже при прохождении через множество слоев, что является особенно важным моментом при наличии шума в данных, где стабильность таких градиентов имеет критическое значение.

2. Сохранение информации. Прямое соединение позволяет беспрепятственно проходить входной информации через сеть, что позволяет сохранить исходные признаки даже в условиях шума.

3. Адаптивная глубина. Сеть имеет возможность выборочно отключать определенные резидуальные блоки (обнуляя их веса), эффективно уменьшая свою глубину в некоторых случаях.

Ансамблевые методы в машинном обучении основаны на принципе объединения множества моделей с целью повышения точности классификации, эффективности и устойчивости [Zhou, 2012]. Данный подход основан на главном, фундаментальном принципе: разнообразие независимых экспертов (моделей) улучшает качество коллективного решения. Ансамблевое обучение является особенно актуальным в контексте рассматриваемой задачи – обучение и классификация в условиях зашумленных данных, поскольку данное разнообразие позволяет нам компенсировать недостатки отдельных моделей, что способствует повышению устойчивости к различным видам шума.

Результат классификации объекта, полученный как объединение предсказаний нескольких классификаторов, формируется как комбинация результатов базовых классификаторов, входящих в ансамбль.

Ансамблевые методы, которые основываются на комбинировании предсказаний нескольких базовых моделей, позволяют существенно улучшить устойчивость и точность классификации зашумленных данных. В особенности это касается адаптивных ансамблевых моделей, учитывающие уверенность и компетентность отдельных моделей в различных областях пространства признаков и динамически адаптирующихся к различного рода шуму.

На основе проведенных исследований был спроектирован программный комплекс, способный эффективно решать задачи классификации зашумленных данных. Ключевой особенностью разработанного программного комплекса является реализация адаптивного ансамбля моделей, который динамически комбинирует результаты классификации нейронной сети и классических алгоритмов машинного обучения в случае, если главный классификатор – нейронная сеть – делает ошибки.

Проектирование адаптивного ансамбля было связано с принятием следующих решений. Ансамбль включает в себя основную нейронную сеть, а также набор вспомогательных моделей, что обеспечивает разнообразие подходов к классификации. В состав ансамбля был включен ряд моделей классификации. Среди них наибольшее влияние на результаты классификации оказали следующие модели:

1. Основная нейронная сеть с резидуальными соединениями.
2. Случайный лес (Random forest).
3. Градиентный бустинг (Gradient boosting).
4. Метод опорных векторов (SVM).
5. К-ближайших соседей (k-NN).

Особенностью ансамбля является наличие механизма адаптивного взвешивания, что позволяет динамически определять вклад каждой модели в итоговое решение на основе её уверенности и исторической точности для конкретного типа данных и шума.

В качестве базовой архитектуры выбрана глубокая нейронная сеть с несколькими скрытыми слоями и резидуальными соединениями. Её особенности:

1. Входной слой с возможностью добавления контролируемого шума.
2. Резидуальные блоки, улучшающие распространение градиентов.
3. Возможность выбора различных функций активации.
4. Многоуровневая регуляризация для предотвращения переобучения.

#### **4. Эксперименты и оценка результатов**

Методика проведения эксперимента с использованием подготовленных программных средств заключается в последовательной реализации таких этапов, как:

- этап подготовки данных;

- этап обучения моделей и оценки их качества;
- этап классификации тестовых примеров при различных уровнях шума в данных. Рассмотрим их подробнее.

Для всестороннего исследования эффективности разработанного программного комплекса были выбраны разносторонние наборы данных с разными характеристиками: размерностью, сбалансированностью классов, сложностью разделяющих поверхностей, областями применения [UCI, 1998]. Каждый набор данных разделялся на обучающую выборку и тестовую выборку.

Каждый тестовый набор данных подвергался воздействию шума различных типов с изменяющимся уровнем интенсивности от 0% до 30% и шагом 5%.

Для каждого набора данных проводилось обучение базовой нейронной сети, и остальных моделей ансамбля.

Для всех сочетаний модели шума и интенсивности шума проводилась классификация тестовых примеров с использованием базовой нейронной сети и остальных моделей ансамбля. Каждый эксперимент повторялся многократно, затем результаты усреднялись.

Для оценки результатов классификации использовалась мера Ассурасы, которая представляет долю правильно классифицированных примеров по сравнению с общим числом примеров тестовой выборки. Также исследовалось влияние на точность классификации тестовых примеров методов предварительной обработки данных, а также использование ансамблевых методов. Результаты оформлялись в виде графиков и таблиц.

Ниже представлены полученные результаты для следующих случаев.

1 Исследование влияния гауссовского шума проводилось на наборе данных Breast Cancer. Этот набор данных был выбран в качестве репрезентативного примера ввиду его сбалансированности и медицинской значимости задачи, где каждый процент точности имеет критическое значение для корректности диагностики.

В табл. 1 представлена зависимость точности классификации тестовых примеров от уровня шума при использовании гауссовской модели шума. Каждое значение в таблице представляет точность классификации тестовых примеров, показанную одним из классификаторов при наличии в данных указанного уровня шума (в процентах). Для каждого уровня шума наилучшие результаты выделены жирным шрифтом.

Как видно из таблицы, разработанная адаптивная ансамблевая модель демонстрирует существенно более высокую устойчивость к гауссовскому шуму по сравнению с отдельно взятыми алгоритмами. Особенно показательна стабильность ансамбля с главенствующей в нем хорошо разработанной нейронной сетью при высоких уровнях шума, где разрыв в точности становится наиболее заметным.

Таблица 1

Зависимость точности классификации тестовых примеров (%) от уровня шума для модели шума «Гауссовский» при работе с набором данных Breast Cancer

Уровень шума %	Ансамблевая модель %	Нейронная сеть %	Random Forest %	Gradient Boosting %	SVM %	K-NN %
0	<b>97</b>	92	96	91	<b>97</b>	94
5	<b>97</b>	91	94	90	<b>97</b>	94
10	<b>96</b>	90	94	90	95	94
15	<b>96</b>	91	94	90	95	95
20	<b>96</b>	91	94	91	<b>96</b>	95
25	<b>95</b>	90	93	91	<b>95</b>	93
30	<b>95</b>	91	92	90	93	93

2. Следующая таблица представляет результаты, полученные для набора данных Wine при наличии импульсного шума. Импульсный шум, как и предполагалось, оказался одним из наиболее разрушительных типов шума и снижает точность классификации тестовых примеров для большинства моделей. Результаты работы с набором данных Wine представлены в табл. 2.

Таблица 2

Зависимость точности классификации тестовых примеров (%) от уровня шума для модели «импульсный шум» при работе с набором данных Wine

Уровень шума %	Ансамблевая модель, %	Нейронная сеть, %	Random Forest, %	Gradient Boosting, %	SVM, %	K-NN, %
0	<b>100</b>	95	95	97	<b>100</b>	95
5	<b>96</b>	81	93	95	80	82
10	<b>93</b>	67	92	90	65	70
15	<b>93</b>	63	89	90	56	67
20	<b>88</b>	55	85	85	53	52
25	<b>89</b>	48	85	86	47	48
30	<b>85</b>	43	82	80	42	43

Разработанный адаптивный ансамбль демонстрирует значительное преимущество над отдельными моделями, особенно при высоких уровнях шума. Так при уровне шума равном 30% точность классификации тестовых примеров у нейронной сети, работающей без поддержки алгоритмов, входящих в ансамбль, падает на более чем 50%. Это значительно превышает аналогичные результаты при гауссовском шуме. Также при высоких уровнях шума адаптивный ансамбль превосходит в среднем на 3% такие алгоритмы, как Random Forest, и Gradient Boosting.



Методы классификации K-NN, SVM и одиночная нейронная сеть показали наименьшую устойчивость к импульсному шуму, что объясняется их чувствительностью к экстремальным значениям.

3. Шум, который проявляется в виде пропущенных значений, представляет собой особую проблему, поскольку многие алгоритмы машинного обучения не способны работать с неполными данными. Эксперименты проводились на наборе данных Heart Disease с последовательным увеличением вероятности появления в данных пропущенных значений от 0 до 30%. Результаты эксперимента на обучающих выборках представлены в табл. 3.

Таблица 3

Зависимость точности классификации тестовых примеров от уровня шума для модели шума «пропущенные значения» при работе с набором данных Heart Disease

Уровень шума %	Ансамблевая модель %	Нейронная сеть %	Random Forest %	Gradient Boosting %	SVM %	K-NN %
0	<b>86</b>	<b>86</b>	80	78	84	80
5	<b>86</b>	<b>86</b>	80	81	85	82
10	<b>85</b>	84	79	79	84	80
15	<b>83</b>	<b>83</b>	77	79	82	81
20	<b>81</b>	79	76	75	<b>81</b>	77
25	<b>82</b>	80	75	74	80	77
30	<b>79</b>	78	74	74	<b>79</b>	77

## 5. Предобработка данных

Важной частью разработанного программного комплекса является *модуль предобработки данных*. Для рассмотренных моделей шума – гауссовский шум, импульсный шум, шум «отсутствие значений» были предложены и программно смоделированы следующие методы обработки.

Известно, что гауссовский шум характеризуется нормальным распределением, что в свою очередь позволяет применять статистически оптимальные методы фильтрации, к ним относится адаптивный фильтр Винера. Данный метод предобработки позволяет снизить уровень шума без потери важной информации, что обеспечивает качество данных для их последующего анализа и классификации. [Gonsales et al., 2018].

Импульсный шум требует специальных методов, способных обнаруживать и корректировать локальные выбросы без размытия значимой информации; к таким методам относится адаптивный медианный фильтр. Метод автоматически подстраивается под локальные характеристики, обеспечивая оптимальный баланс между подавлением шума и сохранением полезной информации. [Hendrycks et al., 2019].

Обработка пропущенных значений и алгоритмы восстановления этих значений рассмотрены в [Fomina et al. 2013]. Прежде всего это метод  $K$  ближайших соседей: пропущенные значения заполняются на основе анализа значений у  $K$  ближайших, наиболее похожих объектов. Сходство объектов определяется с помощью евклидова расстояния в пространстве признаков, не содержащих шума.

Помимо методов и алгоритмов обработки зашумленных данных предложен механизм автоматического определения типа шума на основе статистического анализа.

## Заключение

Проведенные экспериментальные исследования, выполненные с помощью разработанного программного комплекса для классификации зашумленных данных на основе нейронных сетей, показали эффективность предложенного подхода. Предложенный адаптивный ансамбль с главенствующей нейронной сетью продемонстрировал высокую точность классификации для всех исследованных типов шума. Особенно это касается ситуаций, когда уровень шума достигает высоких значений. Экспериментально получены оценки влияния шумов разного типа на успешность классификации.

Разработанный и реализованный механизм ансамблирования успешно расширяет возможности нейронной сети, сохраняя её в качестве основного классификатора и динамически подключая вспомогательные алгоритмы только в случаях, когда нейросеть демонстрирует недостаточную уверенность в своих предсказаниях. Такой подход позволил повысить устойчивость нейросетевого решения к различным типам шума без необходимости кардинального изменения архитектуры самой нейронной сети. Предложенная архитектура нейронной сети с резидуальными соединениями в сочетании с методами предобработки данных и адаптивным ансамблированием, способна обеспечить эффективное решение даже в наиболее сложных случаях.

Реализованные методы предобработки данных, подверженных влиянию шума, способны обеспечить в большинстве случаев значительное повышение точности классификации, прежде всего за счёт сглаживания импульсных выбросов и применения методов восстановления неизвестных значений в данных.

## Список литературы

- [Goodfellow et al., 2016] Goodfellow I., Bengio Y., Courville A. Deep Learning. – Cambridge: MIT Press, 2016.
- [Гафаров и др., 2018] Гафаров Ф.М., Галимьянов А.Ф. Искусственные нейронные сети и приложения: учебное пособие. – Казань: Изд-во Казан. Ун-та, 2018.

- [Еремеев и др., 2021] Еремеев А.П., Варшавский П.Р., Поляков С.А. Программная реализация модуля анализа данных на основе прецедентов для распределенных интеллектуальных систем // Программные продукты и системы. – 2021. – № 34(3).
- [Москвитин и др., 2023] Москвитин В.М., Семёнова Н.И. Влияние шума на рекуррентные нейронные сети с нелинейными нейронами // Известия вузов. Прикладная нелинейная динамика. – 2023. – Т. 31, № 4.
- [Вагин и др., 2008] Вагин В.Н., Головина Е.Ю., Загорянская А.А., Фомина М.В. Достоверный и правдоподобный вывод в интеллектуальных системах / под ред. В.Н. Вагина, Д.А. Поспелова. – 2-е изд. – 2008.
- [Еремеев и др., 2023] Еремеев А.П., Сергеев М.Д., Петров В.С. Интеграция методов обучения с подкреплением и нечеткой логики для интеллектуальных систем реального времени // Программные продукты и системы. – 2023. – № 4.
- [Gonsales et al., 2018] Gonzalez R.C., Woods R.E. Digital Image Processing. – 4th ed. – London: Pearson, 2018.
- [Fomina et al. 2013] Marina Fomina, Alexander Eremeev, Vadim Vagin. Noise models in Inductive Concept Formation // Proceedings of ICEIS 2013, 15th International Conference on Enterprise Information Systems, 2013, Angers – France. – 2013. – Vol. 1. – P. 413-419.
- [He et al., 2016] He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition // Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). – 2016.
- [Zhou, 2012] Zhou Zhi-Hua. Ensemble Methods: Foundations and Algorithms. – Boca Raton: Chapman & Hall/CRC, 2012.
- [UCI, 1998] UCI Repository of Machine Learning Datasets. Available: <http://archive.ics.uci.edu/ml/>.
- [Hendrycks et al., 2019] Hendrycks D., Dietterich T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations // Int. Conf. On Learning Representations (ICLR). – 2019.