Анализ методов и алгоритмов искусственного интеллекта при обработке данных в виде ряда сигналов

П. Ю. Беляев[1], Е. Л. Шейнман[1,2], Ю. В. Ким[1]
[1] *Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)*
[2] *АО «Концерн «Океанприбор»*
Belyaev.edu@gmail.com

***Аннотация*. В работе рассматриваются вопросы повышения точности при выполнении задачи классификации ряда сигналов, где главной задачей является определение класса объекта на основе данных временного ряда данного объекта. Примерами таких сигналов могут быть сигналы ЭКГ, звуки, вибрации и другие. Для успешного решения задачи классификации важно правильно выбрать подходящий метод и качественно подготовить данные для обучения модели, включая предварительную обработку данных и подбор параметров модели. Данное исследование включает в себя обзор методов искусственного интеллекта в области анализа данных на основе ряда сигналов, включая алгоритмы машинного обучения. Для исследования были использованы наборы данных: Sonar, Doppler, Winnipeg. Исходя из сравнения исследуемых методов, наивысшую точность показали SVM, Random Forest, AdaBoost, KNN. Средняя точность классификации составила 0.9.**

Ключевые слова: обработка сигналов; машинное обучение; анализ данных

I. Введение

С развитием технологий и науки о данных искусственный интеллект (ИИ) становится все более распространенным и используется в различных областях, где требуется обработка больших объемов информации. Одной из таких областей является обработка данных в виде ряда сигналов. Ряды сигналов возникают в различных сферах, таких как обработка изображений, звука, временных рядов и других.

Ряд сигналов представляет собой последовательность данных, которая может быть записана в виде временного ряда. В машинном обучении ряды сигналов могут быть использованы для решения различных задач, таких как:

- прогнозирование;
- классификация;
- детектирование аномалий.

Ряд сигналов может быть получен из различных источников, таких как звуковые записи, медицинские измерения, данные с датчиков и многих других. Эти данные могут быть представлены в виде последовательности, где каждый элемент соответствует определенному моменту времени и может быть числовым или категориальным значением. Важным аспектом ряда сигналов является то, что данные в нем не являются независимыми и одинаково распределенными. Например, если мы рассматриваем звуковую запись, то каждый звуковой сигнал зависит от предыдущих и следующих звуковых сигналов, а также от времени, в которое он был записан [1].

Для анализа ряда сигналов в машинном обучении применяются различные методы и алгоритмы, которые позволяют извлекать полезную информацию из данных. Например, методы временных рядов могут быть использованы для выделения повторяющихся паттернов в ряде сигналов, а методы глубокого обучения могут использоваться для распознавания образов и классификации.

В данной статье рассматривается анализ методов и алгоритмов искусственного интеллекта, используемых при обработке данных в виде ряда сигналов. Будут рассмотрены различные подходы, такие как нейронные сети и методы машинного обучения. Также будет проведен анализ применения данных методов в задаче классификации источников при обработке сигнала. В результате будут представлены оптимальные методы для применения различных методов для обработки ряда сигналов.

II. Используемые методы

A. *Описание методов*

Существует множество методов машинного обучения, используемых для решения задач классификации и регрессии. В исследовании ряда сигналов были использованы методы логистической регрессии [2], Random Forest [3], SVM [4], AdaBoost [5], дерево решений [6], Gradient Boosting [7] и KNN [8]. Данные методы являются одними из самых популярных методов машинного обучения, которые используются для решения различных задач.

Логистическая регрессия – моделирует вероятность принадлежности объекта к определенному классу, используя логистическую функцию для преобразования линейной комбинации признаков объекта в вероятность. Другой метод – дерево решений – строит дерево, состоящее из узлов и листьев, каждый из которых соответствует определенному правилу принятия решений на основе значений признаков объекта. SVM (Support Vector Machine) находит гиперплоскость, которая разделяет объекты разных классов в

Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации "Госзадание" №075-01024-21-02 от 29.09.2021 (проект FSEE-2021-0014).

пространстве признаков, максимизируя расстояние между этой гиперплоскостью и объектами каждого класса. AdaBoost (Adaptive Boosting) комбинирует несколько слабых классификаторов в один сильный классификатор, уменьшая ошибку классификации путем взвешивания объектов, которые были неправильно классифицированы на предыдущих итерациях. Random Forest комбинирует несколько деревьев решений, случайным образом выбирая подмножество признаков и объектов для построения каждого дерева, чтобы уменьшить эффект переобучения. Gradient Boosting комбинирует несколько простых моделей, используя градиентный спуск для минимизации ошибки предсказания на каждом шаге. KNN (k-ближайших соседей) использует расстояние между объектами в пространстве признаков для определения класса (или значения для регрессии) нового объекта, находя k ближайших соседей в обучающей выборке.

B. Сравнение методов

Преимущества и недостатки данных подходов представлены в таблице 1.

ТАБЛИЦА I. СРАВНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Метод	Преимущества	Недостатки
KNN	- устойчивость к выбросам - простота реализации -легкая интерпретируемость	- высокие временные и вычислительные затраты
Логистическая регрессия	- простота реализации - низкие временные и вычислительные затраты	- рассчитана на бинарную классификацию - требует дополнительных преобразований для нелинейных функций
Дерево решений	- низкие временные и вычислительные затраты	- склонность к переобучению
SVM	- способность работать с нелинейными функциями	- чувствительность к отсутствующим данным
K-means	- низкие временные и вычислительные затраты	- сложная интерпретируемость - чувствительность к выбросам
AdaBoost	- простота реализации	- высокие временные и вычислительные затраты - сложная интерпретируемость
Random Forest	- способность эффективно обрабатывать данные с большим числом признаков и классов - простота реализации	- высокие вычислительные затраты
Gradient Boosting	- простота реализации	- вычислительные затраты - сложная интерпретируемость

Каждый из этих методов имеет свои преимущества и недостатки, и выбор метода зависит от конкретной задачи и доступных данных. Некоторые методы, такие как SVM и Random Forest, хорошо подходят для работы с большими объемами данных, в то время как другие, такие как многослойный перцептрон, обычно используются для более сложных задач. Комбинация нескольких методов может также улучшить точность модели и повысить ее устойчивость к выбросам и шуму в данных.

III. НАБОР ДАННЫХ

Для исследования были выбраны наборы данных содержащие различные ряды сигналов. Сравнение наборов представлено в табл. 2.

ТАБЛИЦА II. СРАВНЕНИЕ НАБОРОВ ДАННЫХ

Набор данных	Признаки
Sonar data: Набор данных для классификации подводных препятствий	Энергия в заданных 60 полосах частот
Doppler data: Набор данных для распознавания автомобиля, человека, дрона	Признаки матрицы радара (по вертикали – допплеровские частоты, dBm, по горизонтали – ячейки расстояния)
Winnipeg data: Набор данных для классификации пахотных земель	Данные радара, радио-метрическая информация

В наборе данных SonarData содержится ряд данных, на основе которых возможно провести классификацию подводных препятствий на два класса: камень или подводная мина. В качестве признаков набор данных содержит показатели отклика (энергия) для 60 отдельных частот сонара. Набор данных включает в себя 208 наблюдений и 60 признаков. Количество наблюдений для класса «камень» (rock – R) равно 97, для класса «подводная мина» (mine – M) – 111.

Также был использован открытый набор данных DopplerData. Он содержит 17485 записей, из которых 6700 наблюдений для класса «человек», 5065 наблюдений для класса «дрон» и 5720 наблюдений для класса «автомобиль», включающих в себя частотно-модулированный непрерывный сигнал в полосе частот с центром на частоте 8.75 ГГц с максимальной шириной полосы 500 МГц. На основе изначальной матрицы (4092x512) была произведена предобработка для сжатия матрицы до размера найденных объектов (11x61).

Набор данных WinnipegData включает в себя 325834 наблюдений и 7 классов сельскохозяйственных культур. В качестве объектов классификации служат двухвременные данные оптического радара для классификации пахотных земель в табличной форме, полученные из изображений, собранных спутниками RapidEye и беспилотными летательными аппаратами.

IV. РЕЗУЛЬТАТЫ

Была проведена проверка эффективности различных методов машинного обучения на основе собранных данных. Для каждого метода была произведена оптимизация гиперпараметров и произведено обучение. Оценка эффективности методов была произведена на тестовой выборке, с соотношением обучающей и тестовой выборок 80 к 20. Для оценки эффективности использовались метрики Accuracy, Precision, Recall и F1 [10], которые были рассчитаны по (1) – (4) соответственно:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = \frac{2 \cdot (Precision \cdot Recall)}{Precision + Recall} \quad (4)$$

Оценка различных методов производилась на АРМ с процессором Intel Core i7-9700KF и размером оперативной памяти 16 Гб. Поиск оптимальных гиперпараметров для модели проводился с помощью GridSearch и занимал от 5 секунд до 8 часов, обучение модели занимало от 30 секунд до 20 минут – время работы зависело от размера набора данных и количества перебираемых гиперпараметров.

Результаты обучения моделей представлены в табл. 3.

ТАБЛИЦА III. МЕТРИКИ ОЦЕНКИ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ НА ТЕСТОВЫХ ВЫБОРКАХ ДАННЫХ

Метод	**Метрика**	**Sonar data**	**Doppler data**	**Winnipeg data – первый период**	**Winnipeg data – второй период**
Логистическая регрессия	*Accuracy*	0.81	0.89	0.87	0.91
	Precision (macro)	0.81	0.88	0.89	0.9
	Recall	0.79	0.88	0.88	0.91
	F1	0.8	0.88	0.88	0.9
Дерево решений	*Accuracy*	0.69	0.86	0.95	0.95
	Precision (macro)	0.7	0.86	0.94	0.94
	Recall (macro)	0.69	0.86	0.94	0.94
	F1 (macro)	0.69	0.86	0.94	0.94
SVM	*Accuracy*	0.83	0.94	0.95	0.95
	Precision (macro)	0.84	0.94	0.95	0.94
	Recall (macro)	0.81	0.94	0.95	0.95
	F1 (macro)	0.82	0.94	0.95	0.94
AdaBoost	*Accuracy*	0.83	0.9	0.55	0.47
	Precision (macro)	0.83	0.9	0.53	0.51
	Recall (macro)	0.83	0.9	0.63	0.64
	F1 (macro)	0.83	0.9	0.52	0.54
Random Forest	*Accuracy*	0.74	0.85	0.82	0.86
	Precision (macro)	0.74	0.86	0.75	0.72
	Recall (macro)	0.72	0.85	0.59	0.75
	F1 (macro)	0.73	0.85	0.6	0.73
Gradient Boosting	*Accuracy*	0.71	0.93	0.9	0.89
	Precision (macro)	0.75	0.93	0.64	0.67
	Recall (macro)	0.76	0.93	0.65	0.64
	F1 (macro)	0.71	0.93	0.65	0.65
KNN	*Accuracy*	0.64	0.91	0.95	0.96
	Precision (macro)	0.65	0.91	0.94	0.94
	Recall (macro)	0.62	0.91	0.93	0.95
	F1 (macro)	0.62	0.9	0.94	0.95

Исходя из данных таблицы, определено лучшее значение Accuracy для каждого набора данных. Так, лучшие результаты (лучшее значение Accuracy присутствует в двух и более наборах данных для данного метода) для подобранных данных показали методы SVM и KNN. Для методов, показавших лучший результат по метрике Accuracy, были построены матрицы ошибок.

Матрицы ошибок для лучших методов набора данных sonardata представлены на рис. 1.

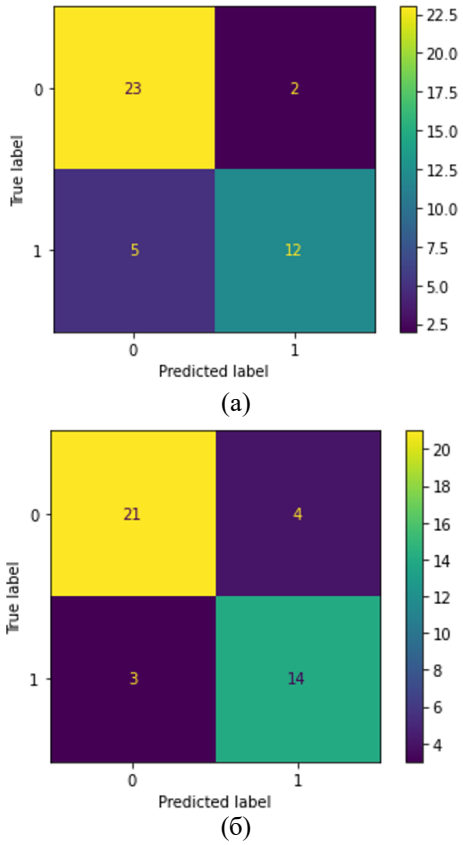


Рис. 1. Матрицы ошибок для sonar data: (а) – SVM; (б) – AdaBoost

Матрицы ошибок для лучших методов doppler data представлены на рис. 2. На рис. 2 можно отметить концентрацию больших значений на главной диагонали матрицы.

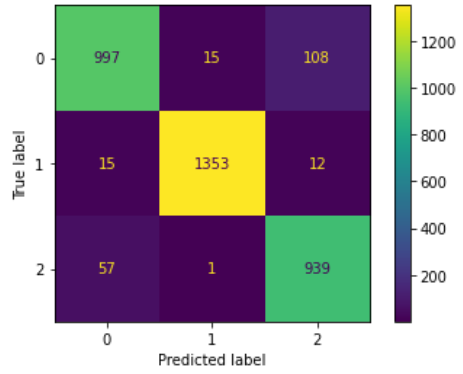
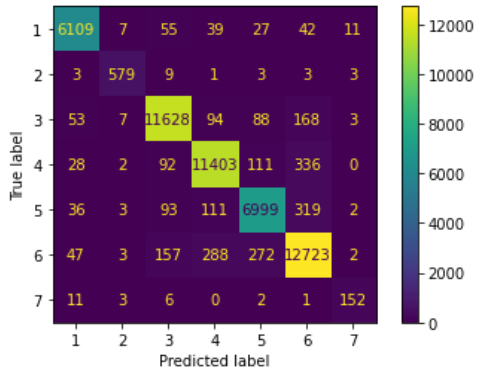
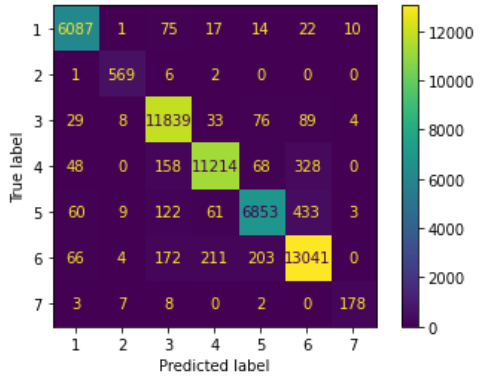


Рис. 2. Матрица ошибок для doppler data (SVM)

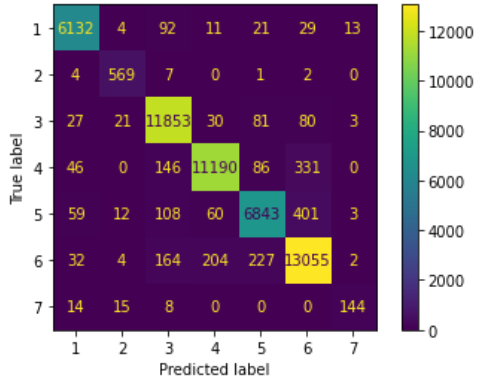
Матрицы ошибок для первого временного периода Winnipeg data представлены на рис. 3.



(а)



(б)



(в)

Рис. 3. Матрицы ошибок для первого временного периода Winnipeg data: (а) – дерево решений; (б) – SVM; (в) – KNN

Матрицы ошибок для второго временного периода Winnipeg data представлены на рис. 4.

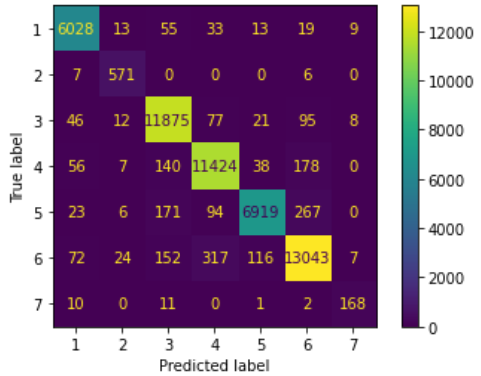


Рис. 4. Матрицы ошибок для второго временного периода Winnipeg data (KNN)

V. Заключение

Таким образом, исходя из совокупности метрик (а именно Accuracy, Precision, Recall, F1), представленных ранее в табл. 3, для работы с наборами входных данных совокупный качественный результат показали следующие алгоритмы машинного обучения:

- для Sonar data – SVM и AdaBoost;
- для Doppler data – SVM;
- для Winnipeg data (первый период) – SVM, KNN;
- для Winnipeg data (второй период) – KNN.

Список литературы

[1] Li Z., Liu R., Wu D. Data-driven smart manufacturing: Tool wear monitoring with audio signals and machine learning // Journal of Manufacturing Processes. 2019. Т. 48. С. 66-76.

[2] PLaValley M. P. Logistic regression // Circulation. 2008. Т. 117. №. 18. С. 2395-2399.

[3] Rigatti S.J. Random forest //Journal of Insurance Medicine. 2017. Т. 47. №. 1. С. 31-39.

[4] Awad M. et al. Support vector machines for classification //Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. 2015. С. 39-66.

[5] Hastie T. et al. Multi-class adaboost //Statistics and its Interface. 2009. Т. 2. №. 3. С. 349-360.

[6] Song Y.Y., Ying L.U. Decision tree methods: applications for classification and prediction // Shanghai archives of psychiatry. 2015. Т. 27. №. 2. С. 130.

[7] Natekin A., Knoll A. Gradient boosting machines, a tutorial //Frontiers in neurorobotics. 2013. Т. 7. С. 21.

[8] Guo G. et al. KNN model-based approach in classification //On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings. – Springer Berlin Heidelberg, 2003. С. 986-996.

[9] Murtagh F. Multilayer perceptrons for classification and regression // Neurocomputing. 1991. Т. 2. №. 5-6. С. 183-197.

[10] Powers D.M.W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation //arXiv preprint arXiv:2010.16061. 2020.