### **МАШИННОЕ ОБУЧЕНИЕ. АЛГОРИТМЫ ДЛЯ ПОТОЧНОЙ ОБРАБОТКИ**

### **СОДЕРЖАНИЕ**

1. ВВЕДЕНИЕ
2. ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ

2.1 Понятие и задачи машинного обучения

2.2 Виды машинного обучения

1. ПОТОЧНАЯ ОБРАБОТКА ДАННЫХ

3.1 Особенности поточных данных

3.2 Проблемы и вызовы поточной обработки

1. АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПОТОЧНОЙ ОБРАБОТКИ

4.1 Инкрементальное обучение

4.2 Онлайн-обучение

4.3 Алгоритмы для обработки потоков данных

1. ПРИМЕНЕНИЕ ПОТОЧНЫХ АЛГОРИТМОВ

5.1 Области применения

5.2 Примеры реализации

1. ЗАКЛЮЧЕНИЕ
2. СПИСОК ЛИТЕРАТУРЫ

### **1. ВВЕДЕНИЕ**

Современный мир генерирует огромные объемы данных ежесекундно. Социальные сети, финансовые транзакции, датчики интернета вещей (IoT) и другие источники создают непрерывные потоки информации, которые необходимо анализировать в реальном времени. Традиционные методы машинного обучения, требующие наличия полного набора данных перед обучением, становятся неэффективными в таких условиях.

Поточная обработка данных (stream processing) позволяет решать эту проблему, обрабатывая информацию по мере ее поступления. Алгоритмы машинного обучения для поточной обработки способны обучаться на лету, адаптируясь к изменениям в данных, что делает их незаменимыми в таких областях, как:

* Финансовый мониторинг (обнаружение мошенничества)
* Анализ сетевого трафика (кибербезопасность)
* Персонализированные рекомендации
* Прогнозирование в реальном времени

Цель данного реферата — рассмотреть ключевые алгоритмы машинного обучения, применяемые для поточной обработки данных, их преимущества, ограничения и практическое применение.

### **2. ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ**

#### **2.1 Понятие и задачи машинного обучения**

Машинное обучение (Machine Learning, ML) — это подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться на данных. В отличие от традиционного программирования, где правила обработки данных задаются явно, ML-алгоритмы автоматически выявляют закономерности и принимают решения на основе предыдущего опыта.

Основные задачи машинного обучения:

1. **Классификация** — отнесение объекта к одному из заранее определенных классов (например, распознавание спама в электронной почте).
2. **Регрессия** — предсказание числовых значений (например, прогнозирование цен на акции).
3. **Кластеризация** — группировка объектов без предварительных меток (например, сегментация клиентов).
4. **Обнаружение аномалий** — выявление нестандартных событий (например, мошеннических транзакций).

#### **2.2 Виды машинного обучения**

1. **Обучение с учителем (Supervised Learning)**
   1. Используется размеченный набор данных (имеются входные данные и правильные ответы).
   2. Примеры алгоритмов:
      1. Линейная регрессия
      2. Метод опорных векторов (SVM)
      3. Деревья решений
2. **Обучение без учителя (Unsupervised Learning)**
   1. Данные не размечены, алгоритм ищет скрытые структуры.
   2. Примеры алгоритмов:
      1. K-средних (K-Means)
      2. Иерархическая кластеризация
3. **Обучение с подкреплением (Reinforcement Learning)**
   1. Алгоритм учится на взаимодействии со средой, получая "вознаграждение" за правильные действия.
   2. Применение: робототехника, игры (AlphaGo).

### **3. ПОТОЧНАЯ ОБРАБОТКА ДАННЫХ**

#### **3.1 Особенности поточных данных**

Поточные данные обладают следующими характеристиками:

* **Непрерывность** — данные поступают постоянно (например, показания датчиков).
* **Высокая скорость** — необходимость обработки в реальном времени.
* **Ограниченный объем памяти** — невозможно хранить весь поток.
* **Изменчивость** — статистические свойства данных могут меняться со временем (Concept Drift).

#### **3.2 Проблемы и вызовы поточной обработки**

1. **Дрейф данных (Concept Drift)**
   1. Пример: в финансовых данных мошенники меняют стратегии, и модель должна адаптироваться.
   2. Решение: алгоритмы с динамическим переобучением.
2. **Ограниченные ресурсы**
   1. Нельзя хранить весь исторический поток.
   2. Решение: инкрементальные методы, сэмплирование.
3. **Задержки обработки**
   1. Необходимость баланса между скоростью и точностью.

### **4. АЛГОРИТМЫ ДЛЯ ПОТОЧНОЙ ОБРАБОТКИ**

#### **4.1 Инкрементальное обучение**

Алгоритмы, которые обновляют модель по мере поступления новых данных:

* **Stochastic Gradient Descent (SGD)**
  + Обновляет веса модели на каждом новом примере.
  + Применение: линейная регрессия, нейронные сети.
* **Perceptron**
  + Простейшая модель для бинарной классификации.

#### **4.2 Онлайн-обучение**

Методы, обрабатывающие данные по одному примеру:

* **Hoeffding Trees**
  + Решающие деревья для потоков данных.
  + Используют статистические границы для выбора разделений.
* **Online Random Forests**
  + Ансамбль Hoeffding Trees для повышения точности.

#### **4.3 Алгоритмы для обработки потоков**

* **K-Means (мини-пакетная версия)**
  + Кластеризация в реальном времени.
* **Half-Space Trees**
  + Обнаружение аномалий в потоках.

### **5. ПРИМЕНЕНИЕ ПОТОЧНЫХ АЛГОРИТМОВ**

#### **5.1 Области применения**

1. **Финансы**
   1. Обнаружение мошенничества в транзакциях.
2. **IoT**
   1. Мониторинг оборудования в реальном времени.
3. **Рекомендательные системы**
   1. Адаптация рекомендаций под текущие предпочтения пользователя.

#### **5.2 Примеры реализации**

* **Apache Kafka + Spark MLlib**
  + Обработка потоков данных с использованием ML-алгоритмов.
* **TensorFlow Extended (TFX)**
  + Потоковое обучение нейросетей.

### **6. ЗАКЛЮЧЕНИЕ**

Алгоритмы машинного обучения для поточной обработки данных позволяют эффективно работать с непрерывными потоками информации, адаптируясь к изменениям в реальном времени. Их применение критически важно в таких областях, как финансы, IoT и кибербезопасность. Основные вызовы — дрейф данных и ограниченные ресурсы — требуют дальнейшего развития методов онлайн-обучения и оптимизации вычислений.

### **7. СПИСОК ЛИТЕРАТУРЫ**

1. Bishop C. M. Pattern Recognition and Machine Learning. – Springer, 2006.
2. Gama J. Knowledge Discovery from Data Streams. – Chapman & Hall, 2010.
3. Goodfellow I., Bengio Y., Courville A. Deep Learning. – MIT Press, 2016.
4. Scikit-learn Documentation. – URL: [https://scikit-learn.org](https://scikit-learn.org/)
5. Apache Spark MLlib Guide. – URL: <https://spark.apache.org/docs/latest/ml-guide.html>