

ФОРМИРОВАНИЕ НАБОРА ДАННЫХ ДЛЯ ОБУЧЕНИЯ МЕТОДА СЛУЧАЙНЫХ ЛЕСОВ В ЗАДАЧАХ РАСПОЗНАВАНИЯ ЛЖИ ПО МИМИКЕ ЛИЦА

Магистрант, Котова К. А.

Научный руководитель – кандидат технических наук Садов В. С.

Белорусский государственный университет

Кафедра интеллектуальных систем

Минск, Беларусь

Введение

Распознавание лжи является актуальной задачей в психологии, криминалистике и информационной безопасности. Современные исследования направлены на разработку автоматизированных систем анализа, использующих методы машинного обучения для выявления признаков лжи по выражению лица, мимике и микровыражениям.

Одним из эффективных алгоритмов для решения задачи классификации является метод случайных лесов (Random Forest, RF). Этот метод позволяет выявлять закономерности в данных, анализировать важность признаков и строить надежные модели. Однако для успешного обучения модели необходимо тщательно подготовить и сбалансировать набор данных, содержащий изображения высокого качества и видеофрагменты с выражениями лица во время произнесения правды и лжи, конкретного человека.

Цель: разработка и анализ набора данных для обучения модели случайного леса, предназначенного для распознавания лжи при коммуникации людей. В статье описываются принципы работы алгоритма случайных лесов и процесс формирования набора данных, включая его структуру и особенности разметки.

Задачи: изучить алгоритм метода машинного обучения Random Forest, разработать наборы данных конкретного человека: изображения для распознавания эмоций и видеоизображения для извлечения микровыражений.

Метод случайных лесов

Метод случайных лесов является алгоритмом машинного обучения с учителем, который применяется для задач классификации и регрессии.[1] Он основан на построении ансамбля деревьев решений, каждое из которых обучается на случайной подвыборке данных. Для получения окончательного результата применяется механизм голосования: прогнозы всех деревьев объединяются, и выбирается наиболее вероятный класс. Процесс классификации с помощью алгоритма случайных лесов представлен на рисунке 1.

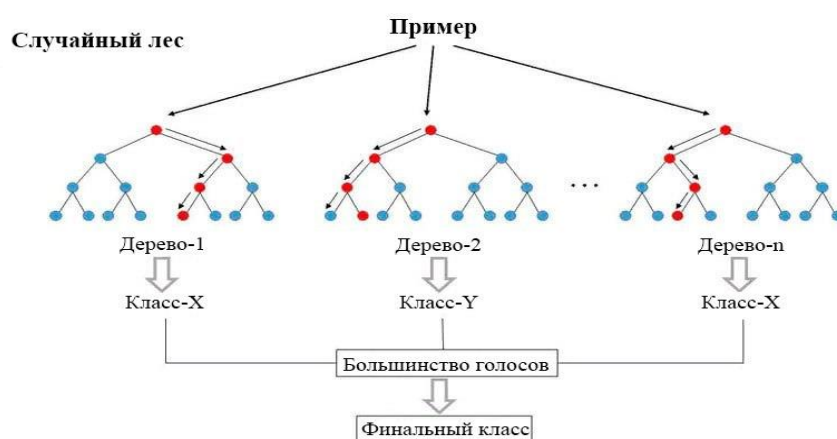


Рисунок. 1. Алгоритм множественной классификации с помощью метода случайного леса

Алгоритм случайных лесов имеет следующие преимущества:

- Способность работать с высокоразмерными данными;
- Устойчивость к переобучению благодаря усреднению решений нескольких деревьев;
- Высокая точность классификации по сравнению с отдельными деревьями решений.

Для построения различных деревьев в рамках одного случайного леса используются два подхода:

1. Случайное разделение обучающей выборки, что позволяет моделям анализировать различные подмножества данных.
2. Случайная выборка из набора признаков, что снижает корреляцию между деревьями и повышает устойчивость модели.

В контексте распознавания лжи метод случайных лесов позволяет анализировать различные признаки мимики, микровыражений и других поведенческих характеристик, выделяя закономерности, характерные для правдивых и ложных высказываний.

Формирование набора данных

Для обучения модели случайного леса необходимо подготовить специализированный качественный набор данных, содержащий изображения лица и видеозаписи с примерами произнесения правды и лжи. [2]

Набор данных состоит из двух частей:

1. Изображения лица с различными эмоциями (рисунок 2)

- 125 изображений конкретного человека.
- Для каждой эмоции в тренировочном датасете содержится 14 изображений, а в тестовом – по 4.
- Изображения сбалансированы таким образом, чтобы равномерно представить различные выражения лица.

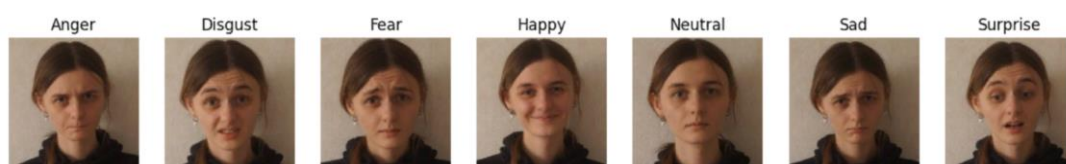


Рисунок 2. Пример изображений из первой части набора данных

2. Видеозаписи произнесения правды и лжи

- 12 видеофайлов, содержащих примеры произнесения лжи.
- 10 видеофайлов, содержащих примеры произнесения правды.
- Видеозаписи анализируются и извлекаются микровыражения.

На рисунке 3 продемонстрировано кратковременное выражение лица, возникшее сразу же после ответа на вопрос. На изображении видны поджатые и напряженные губы. Вероятно, происходит попытка скрыть улыбку или за улыбкой спрятать настоящую эмоцию.



Рисунок 3. Демонстрация микровыражения из второй части набора данных

Для корректного обучения модели данные были размечены следующим образом:

- Видеоразметка включает временные метки ключевых моментов высказываний.
- Изображения аннотированы с указанием наличия микровыражений.
- Выделены характерные особенности мимики, такие как поджатые губы, напряжение лицевых мышц, изменение направления взгляда.

Алгоритм распознавания лжи

Алгоритм распознавания лжи по эмоциям с микровыражениями (рис. 4) заключается в следующем: обучение метода классификации эмоций по ключевым точкам лица из изображений набора данных различных эмоций; извлечение микровыражений из набора данных видеоизображений; предсказание эмоций на основе ключевых точек лица микровыражений; анализ полученных результатов и формирование математического выражения лжи конкретного человека.

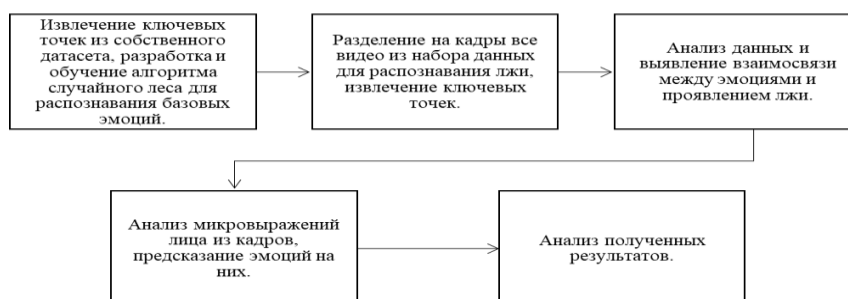


Рисунок 4. Алгоритм исследования микровыражений в общем виде

Заключение

В данной статье рассмотрены особенности формирования набора данных для обучения модели случайного леса, предназначенной для распознавания лжи. Описаны структура датасета, принципы его балансировки и разметки. Метод случайных лесов выбран в качестве алгоритма классификации, так как он демонстрирует высокую точность и устойчивость к переобучению.

Будущие исследования направлены на расширение набора данных, создание автоматизированной системы извлечения микровыражений, а также тестирование различных методов предобработки изображений для улучшения качества классификации.

Литература

1. Головатая Е.А., Нейросетевые технологии в обработке и защите данных: учеб. пособие / Е. А. Головатая, А. В. Курочкин. – БГУ, 2021. – 151 с.
2. Котова, К.А. Алгоритм анализа микровыражений лица при выявлении лжи/К.А. Котова, В.С. Садов//Компьютерные технологии и анализ данных (СТДА'2024): материалы IV Междунар. науч.-практ. конф., Минск, 25–26 апр. 2024 г. / Белорус. гос. ун-т; редкол.: В. В. Скакун (отв. ред.) [и др.]. – Минск: БГУ, 2024. – С.- 68-71.