Слайд 1

Здравствуйте, уважаемые члена комиссии. Меня зовут Султанов Азамат.

Тема моего исследования - Обнаружение внутреннего нарушителя путём выявления стрессового состояния пользователя

Слайд 2,3

В ходе анализа предметной области было найдено несколько работ, в которых авторы показали, что выявление стресса с использованием различных биометрических показателей позволяет обнаруживать внутреннего нарушителя. В данных работах биометрические показатели отслеживались с помощью специального дорогостоящего оборудования, которое зачастую очень сложно найти, например, в обычном офисе или дома, что является причиной усложнённой интеграции исследованных методов на практике.

Однако есть метод, который интересен именно с точки зрения перспективы интеграции, в силу доступности датчиков, в роли которых выступают клавиатура и мышь. К сожалению, именно в данной области найдена всего лишь одна работа, где в качестве биометрических показателей используется динамика взаимодействия с клавиатурой и мышью. Именно это побудило меня на развитие и усовершенствование идеи использования клавиатуры и мыши для детекции стресса, поэтому целью моей работы стало оценить возможность выявления стрессового состояния пользователя на основе анализа взаимодействия с клавиатурой и мышью.

Слайд 4

В силу отсутствия датасетов в исследуемой области, возникла необходимость организации процесса сбора данных.

Для этого были предложены сценарии, спроектированные таким образом, чтобы максимально близко описать ситуации, которые могли бы произойти в реальности. Часть этих сценариев описывает случаи правомерного поведения сотрудника, которые не сопровождаются стрессом, другая описывает ситуации, в которых имитируются действия внутреннего нарушителя, предполагающие индукцию стресса в силу наложенных временных ограничений.

Для сбора данных было написано специальное ПО, фиксирующее в фоновом режиме события мыши и клавиатуры в момент выполнения сценариев участниками эксперимента.

Слайд 5

Так как анализ данных в работе основан на применении алгоритмов машинного обучения, то дальнейшим шагом стало выделение признаков из файлов логирования. Были вычислены, как временные, так и частотные признаки.

В результате выделения признаков был получен датасет на основе которого производилось дальнейшее исследование.

Слайд 6

Перед началом обучения моделей были проделаны шаги предобработки данных, включающие удаление признаков, связанных с редкими событиями, удаление признаков с маленьким значением стандартного отклонения и заполнение пустот в датасете медианами соответствующих признаков.

Слайд 7

На текущих графиках приведены усреднённые значения признаков, которые были по отдельности рассчитаны для категорий нормального и аномального поведений.

Красными зонами выделены наиболее информативные признаки. Они считаются наиболее информативными, потому что именно в этих зонах наблюдается существенное различие значений для нормального и аномального поведений.

Слайд 8

Для того, чтобы отобрать наиболее информативные признаки, тем самым уменьшить размерность пространства признаков, был применён алгоритм отбора K лучших признаков на основе критерия хи-квадрат. Размерность пространства признаков уменьшилась до 3-х. Отобранные признаки выделены зелёным цветом. Смотря на график, можно сделать вывод о том, что отобраны именно те признаки, в которых значения для двух классов сильно отличаются, что действительно говорит о значимости данных признаков.

Нулевая гипотеза — принимаемое по умолчанию предположение о том, что не существует связи между двумя наблюдаемыми событиями, феноменами.

Слайд 9

Для того, чтобы ещё раз убедиться в этом было построено распределение примеров датасета в пространстве признаков. На данном графике, представленном с различных ракурсов, видно, что классы сгруппированы в различных областях пространства и не пересекаются.

Слайд 10

С использованием итогового набора признаков были обучены модели бинарных классификаторов на основе различных алгоритмов машинного обучения. Получены следующие результаты. Лучше всех себя показала модель на основе алгоритма случайного леса.

Слайд 11

На данных графиках можно наблюдать результаты классификации лучшего классификатора на обучающей и тестовой выборках с разных ракурсов.

Слайд 12

Также были обучены модели обнаружения аномалий. Данные модели получали на вход примеры только с нормальным поведением. Наилучшей оказалась модель на основе алгоритма изолирующего леса.

Слайд 13

Результаты этой модели как для обучающей, так и для тестовой выборки можно наблюдать на данных графиках. Также с различных ракурсов.

Слайд 14

В ходе проведённого исследования были:

- Проанализированы существующие методы обнаружения внутреннего нарушителя с использованием биометрических показателей на основе алгоритмов машинного обучения
- Реализованы процессы накопления данных, предобработки данных, обучения и оценки моделей классификаторов и моделей обнаружения аномалий
- Наилучшие результаты получены для моделей на основе алгоритмов случайного леса (Точность - 88%) и изолирующего леса (Точность – 100%)