**Abstract.**

Мы исследовали семейство отравляющих атак против Метода опороных векторов(SVM). Эо атаки вводятся специально в созданной трейн, чтобы увеличить тестовую ошибку SVM. Ключевая мотивация для этих атак - это факт того, что большинство обучащих алгоритов предполагают что их трейн получен из натурального или грамотно-подобрного распределения. Однако, это предположение обычео не выполняется в установках чувстительных к безопасности. Как мы покажем, аккуратное влияние может, в какой то степени, предсказать изменение предсказательной функции из-за вредоносного вода и использовать это возможность для создания вредоносной даты.

Предложеная атака использует стратегию градиентного подьеам в который градиент вычисляется на харатеристиках оптимального решеня SVM. Этот метод может быть введен в ядро и позвоялет сконструировать атаку во входящем просранстве даже в не-линеном ядре. Мы экспериментально показали, что наша процедура увеличения градиента с большой вероятностью находит зороший локальный максимум невыпуклой поверхности валидационной ошибки, которая существуенно увеличивает тестовую ошибку классификатора.

**1. Introduction**

ML техники стремительно появились как ждый инструмент в разнообразии онлайн больших систем приложений, потому что они могут выявлять спрятанные закономерности в больших сложных наборах данных, адаптируясь к новому поведению, и обеспечивать статистическую обоснованность для процесса принятия решений. Разработчики приложений могут использовать обучение таким образом, чтобы помогать решать задачи, который называются big-data problembs и они включает число связанных с безопасностью задач, в частности фокусируясь на нахождении вредоносного или нетепичного поведения. Фактически, подходы обучения уже были использованы или предлогалаись как решения задач, связанных с безопасноть., включая спам, червей вторжения и мошенничества. К несчастью, в этих областях, дата это обычно не только нестанционарное но и также имеет уязвимый компонент, и гибкость предоставленния техникам обучения может использована атакующих для достижения его целей. Например, в спам-детекции, нападающие обычно адаптируют их подходы основаные на популярных спам детекторах, и часто умный нападающий изменит его поведение или увильнет, либо введет в заблуждение.

В ответ на угрозу недобросовестного манипулирования данными несколько предложенных методов обучения явно учитывают определенные типы поврежденных данных. Атак против обучающих алгоритмов могут быть классифицированы, среди других категорий, в *причинные ( манипуляция над тренировочной информацией) и разведочный(ииследования классификатора). Отравление* относится к причинным атакам в который специально созданые атаки вводяться в тренировочный датасет. Это атаки особенно важны с практической точки зрения , так как атакующих обычно не имеет прямого доступа к существующему трейну, но может предложить новый обучащие данные, т. е. Открытые репозитории и соблазн часто собирают вредоносные примеры для обучения, которые предоставляют возможность нападающему для отравления тренировочной даты. Отравляющие атаки могут быть раньше были изучеены для для простых методов детекции аномалий.

В этой статье, мы проверим семейтво отравляющих атак против SVM. Следуя обычному методу анализа защиту для ML, мы допустим, что атакующий знает алгоритм обучения и может взять информацию из основного распределения информации. Далее, мы предположим, что наш атакующий знает трейн испольщующися обучаемым, обычно, в нереалистичном допущении, но в реальном мире, злоумышленник мог бы вместо этого использовать суррогатный

обучающий набор, взятый из того же дистрибутива и наш подход анализ возможностей злоумышленника в худшем случае. Под этими упущениями, мы презентуем новый метод, который атакующий может использовать для создания точки данных, которая существенно уменьштает очность SVM классификатора.

Предложенный метод основан на свойствах оптимального решения задачи SVM обучения. Как было впервые показано в методике поэтапного обучения, это решение плано зависит от параметров задачи квадратичного программирования и геометрии дата точек. Мы покажем, что такие найдены точки могут быть сформулирована как оптиимзазция по показателю производительности, при усвлоии сохранения оптимального задачи обучения SVM. Хотя поверхность тестируемой обычно невыпуклая, процедура подьема по градиенту в нашем методу надежно назеодет локальный максимум поверхности тестируемой ошибки.

Данный предложенный метод зависит только от градиентов скалярныхх произведений между точками во входном пространстве, и следовательно, можт быть ядризован. Это отличается от предыдущей работы, свзяанной с созданием специальных точек атаки, в которых атака может только быть сконструирована в пространстве признаков для нелинейного случая. Последнее это сильное препятствие для атакующего, так как он должуен конструировать дату во входном простнатсве и не имеет практического средства доступа к пространству признаков. Следовательно, предложенный метод ломает новую почву в улучшении влияния основанных на данных атак против ядерных обучающих алгоритмов и подчеркивает необходимость рассмотрения сопротивления против негативной обучающей даты как важдый фактор конструирования обучащих алгоритмов.