УДК 004.056.53

А.Н. НОСКОВ, А.А. ЧЕЧУЛИН, Д.А. ТАРАСОВА

ИССЛЕДОВАНИЕ ЭВРИСТИЧЕСКИХ ПОДХОДОВ К ОБНАРУЖЕНИЮ АТАК НА ТЕЛЕКОММУНИКАЦИОННЫЕ СЕТИ НА БАЗЕ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Носков А.Н., Чечулин А.А., Тарасова Д.А. **Исследование эвристических подходов к обнаружению атак на телекоммуникационные сети на базе методов интеллектуального анализа данных.**
Аннотация. Анализ методик систем обнаружения сетевых атак является перспективным направлением в области защиты сетей и сетевых систем. В статье рассматривается подход к оценке алгоритмов и механизмов обнаружения атак. Новизна предлагаемой методики заключается в возможности создания самообучающихся систем для обнаружения вторжения. В статье рассмотрены основные элементы алгоритмов обнаружения атак.
Ключевые слова: анализ методов обнаружения атак, системы обнаружения вторжения, нежелательный трафик, метод опорных векторов.

Noskov A.N., Chechulin A.A., Tarasova D.A. **Investigation of Heuristic Approach to Attacks on the Telecommunications Network Detection based on Data Mining Techniques.**
Abstract. Analysis of Intrusion Detection System techniques is a perspective area for the protection of networks and network systems. This paper presence the overview of attack detection mechanisms based on data mining approach. The novelty of the this kind of mechanisms is the ability to create self-learning systems for intrusion detection. Also the article describes the basic elements of intrusion detection algorithms.
Keywords: analysis of methods of intrusion detection, intrusion detection systems, malicious traffic, Support Vector Machines.

1. Введение. Одной из фундаментальных научных проблем является разработка методологических основ обеспечения безопасности сетей и обнаружения нежелательного трафика. Особенно важный фактор - время реакции системы на подозрительный трафик. Разработка такой системы обнаружения атак является одной из задач, направленной на повышение защищенности телекоммуникационных сетей. Данные системы используются не только в частных или коммерческих целях, но и в критически важных инфраструктурах государственного значения, вывод из строя которых в результате успешно выполненной атаки может привести к значительным материальным, финансовым и даже политическим потерям.

Существует большое количество исследований в области обнаружения атак, направленных на информационно-телекоммуникационные системы. При этом современные методы обнаружения вторжений базируются на двух принципах: сигнатурный (формальный) и эвристический (обнаружение аномалий, базирующееся на моделях штатного функционирования наблюдаемой

информационной системы) [1]. Однако, существующие методики, как правило, или характеризуются большим количеством пропущенных атак (сигнатурные методы) или высокими требованиями к доступным вычислительным и временным ресурсам (эвристические методы). К тому же, существующие методики обычно ограничиваются детальным исследованием только части характеристик процессов, происходящих в сети, что также приводит к понижению качества выявления атак [21].

Кроме теоретических подходов, для обеспечения защищенности компьютерных сетей были созданы системы, которые классифицируют сетевую активность различных программ. В случае совпадения с ситуацией, определенной экспертом, такие системы предлагают пользователю прекратить действия возможно вредоносного ПО и нейтрализовать произведенные им изменения. Однако большинство современных систем сетевой безопасности не имеют возможности самообучения и оперируют только заложенными в них вручную правилами [7].

Например, в Приказе Федеральной службы по техническому и экспортному контролю (ФСТЭК России) от 11 февраля 2013 г. N 17 г. Москва, Методический документ “Меры защиты информации в государственных информационных системах” (утв. Федеральной службой по техническому и экспортному контролю 11 февраля 2014 г.) представлены методы и рекомендации по использованию систем СОВ для обеспечения безопасности сети и данных государственных и не государственных предприятий.

Сложность таких систем, большое количество используемых программно-аппаратных средств, множество различных событий безопасности усложняют процесс обнаружения вредоносной активности. Поэтому задача повышения защищенности информационно-телекоммуникационные систем от различных угроз является важной фундаментальной проблемой [8].

В рамках данной работы предполагается разработать новый подход, в основу которого ляжет комплексная система анализа, основанная на методах интеллектуальной оценки данных, позволяющая учесть различные характеристики вторжения и принять эффективное решение на основе его глубокого анализа. Кроме того, предлагаемый подход позволит выявить вредоносную сетевую активность, использующую новые уязвимости. Применение методов искусственного интеллекта (ИИ) позволит ввести в системы защиты свойство самообучения и обеспечит обнаружение угроз «на лету». В качестве источников исходных данных могут выступать как аппаратные средства сетевой инфраструктуры, так и распределенные системы управления информацией и событиями безопасности (Security

Information and Events Management, SIEM), активно развивающиеся в последние годы. Таким образом, разработанный подход к обнаружению атак на основе методов интеллектуального анализа данных, позволит повысить эффективность отслеживания ситуации по безопасности и поддержки принятия решений в информационно-телекоммуникационных системах.

2. Исследование существующих алгоритмов обнаружения вторжений. В настоящее время существует множество различных видов сетевых атак, которые используют как уязвимости операционной системы, так и иного установленного программного обеспечения системного и прикладного характера. Злоумышленники постоянно совершенствуют методы нападения, результатом которых может являться кража конфиденциальной информации, выведение системы из строя, либо ее полный «захват» с последующим использованием, как части зомби-сети для совершения новых атак [5].

Для того чтобы своевременно обеспечить безопасность компьютера, важно знать, какого рода сетевые атаки могут угрожать ему. Известные сетевые угрозы можно условно разделить на три большие группы [14]:

Сбор информации – этот вид угроз сам по себе не является атакой, а обычно предшествует ей, поскольку является одним из основных способов получить сведения об удаленном компьютере. Этот способ заключается в сканировании хостов или отдельных UDP/TCP-портов, используемых сетевыми сервисами на интересующем компьютере, для выяснения их состояния (закрытые или открытые порты). Сканирование портов позволяет понять, какие типы атак на данную систему могут оказаться удачными, а какие нет. Кроме того, полученная в результате сканирования информация («слепок» системы) даст представление злоумышленнику о типе операционной системы на удаленном компьютере. А это, в свою очередь, еще сильнее ограничивает круг потенциальных атак и, соответственно, время, затрачиваемое на их проведение, а также позволяет использовать специфические для данной операционной системы уязвимости.

DoS-атаки или атаки, вызывающие отказ в обслуживании – это атаки, результатом которых является приведение атакуемой системы в нестабильное, либо полностью нерабочее состояние. Последствиями такого типа атак могут стать повреждение или разрушение информационных ресурсов, на которые они направлены, и, следовательно, невозможность их использования.

Существует два основных типа DoS атак:

– отправка компьютеру-жертве специально сформированных пакетов, не ожидаемых этим компьютером, что приводит к перезагрузке или остановке системы;

– отправка компьютеру-жертве большого количества пакетов в единицу времени, которые этот компьютер не в состоянии обработать, что приводит к исчерпанию ресурсов системы [6].

Атаки-вторжения, целью которых является «захват» системы. Это самый опасный тип атак, поскольку в случае успешного выполнения, система оказывается полностью скомпрометированной перед злоумышленником. Данный тип атак применяется, когда необходимо получить конфиденциальную информацию с удаленного компьютера (например, номера кредитных карт, пароли) либо просто закрепиться в системе для последующего использования ее вычислительных ресурсов в целях злоумышленника (использование захваченной системы в зомби-сетях), либо как плацдарм для новых атак.

Данная группа является также самой большой по количеству включенных в нее атак. Их можно разделить на три подгруппы в зависимости от операционной системы: атаки на Microsoft Windows-системы, атаки на Unix-системы, а также общая группа для сетевых сервисов, использующихся в обеих операционных системах.

Наиболее распространенными видами атак, использующих сетевые сервисы операционной системы, являются:

– атаки на переполнение буфера;

– атаки, основанные на ошибках обработки данных.

Преимущество сетевых систем обнаружения вторжений (ССОВ, NIPS) заключается в том, что они способны обнаруживать как угрозы, приходящие извне, так и угрозы, исходящие из локальной сети. Для этого, топология сети выстраивается таким образом, чтобы весь сетевой трафик проходил через анализаторы ССОВ (см. рисунок 1).

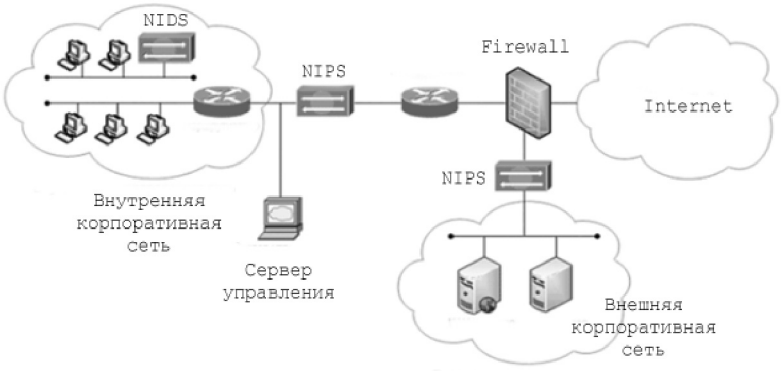

Рис. 1 Структура защищенной компьютерной сети

Обнаружение атак на сетевом уровне сводится к двум методам:

– *Сигнатурный метод* сводится к поиску признаков уже известных атак. Преимущество сигнатурного метода в том, что он практически не подвержен ложным срабатываниям. Недостатком этого метода является невозможность обнаруживать незаложенные в систему атаки.

– *Метод поиска аномалий* (эвристический) позволяет реагировать на ранее неизвестные атаки, но подвержен ложным срабатываниям и требует точной настройки для каждого наблюдаемого объекта [9].

3. Использование машинного обучения в задаче обнаружения сетевых угроз. Существует несколько основных алгоритмов для построения и обучения моделей классификаторов, применимых для решения задачи обнаружения сетевых атак. Рассмотрим некоторые из них более подробно.

3.1. Динамические байесовские сети. Байесовская сеть (БС) — это графическая вероятностная модель, представляющая собой множество переменных и их вероятностных зависимостей. БС представляют собой удобный инструмент для описания достаточно сложных процессов и событий с неопределенностями. БС оказалась особенно полезной при разработке и анализе машинных алгоритмов обучения. Основной идеей построения графической модели является понятие модульности, то есть разложение сложной системы на простые элементы. Для объединения отдельных элементов в систему используются результаты теории вероятностей.

Динамические байесовские сети являются обобщенной моделью в пространстве состояний. Название «динамические» указывает не на зависимость структуры от времени, а только на зависимость от моделирования процесса.

Достоинства байесовских сетей:

– в модели определяются зависимости между всеми переменными, это позволяет легко обрабатывать ситуации, в которых значения некоторых переменных неизвестны;

– байесовские сети достаточно просто интерпретируются и позволяют на этапе прогностического моделирования легко проводить анализ по сценарию "что, если";

– байесовский метод позволяет естественным образом совмещать закономерности, выведенные из данных, и, например, экспертные знания, полученные в явном виде;

– использование байесовских сетей позволяет избежать проблемы переучивания (overfitting), то есть избыточного усложнения

модели, что является слабой стороной многих методов (например, деревьев решений и нейронных сетей).

Байесовский подход имеет следующие недостатки:

– перемножать условные вероятности корректно только тогда, когда все входные переменные действительно статистически независимы, хотя часто данный метод показывает достаточно хорошие результаты при несоблюдении условия статистической независимости, но теоретически такая ситуация должна обрабатываться более сложными методами, основанными на обучении байесовских сетей [10];

– невозможна непосредственная обработка непрерывных переменных – требуется их преобразование к интервальной шкале, чтобы атрибуты были дискретными, однако, такие преобразования иногда могут приводить к потере значимых закономерностей;

– на результат в байесовском подходе влияют только индивидуальные значения входных переменных, комбинированное влияние пар или троек значений разных атрибутов здесь не учитывается [11]. Это могло бы улучшить качество классификационной модели с точки зрения ее прогнозирующей точности, однако, увеличило бы и количество проверяемых вариантов.

3.2. Метод k ближайших соседей. Следует сразу отметить, что метод "ближайших соседей" ("nearest neighbour") относится к классу методов, работа которых основывается на хранении данных в памяти для сравнения с новыми элементами. При появлении новой записи для прогнозирования находятся отклонения между этой записью и подобными наборами данных, и наиболее подобная (или ближайший сосед) идентифицируется.

При таком подходе используется термин "k-ближайших соседей" ("k-nearest neighbour"). Термин означает, что выбирается k "верхних" (ближайших) соседей для их рассмотрения в качестве множества "ближайших соседей". Поскольку не всегда удобно хранить все данные, иногда хранится только множество "типичных" случаев. В таком случае используемый метод называют рассуждением по аналогии (Case Based Reasoning, CBR), рассуждением на основе аналогичных случаев, рассуждением по прецедентам [12].

Преимущества метода:

– Простота использования полученных результатов;

– Решения не уникальны для конкретной ситуации, их использование возможно для других случаев;

– Целью поиска является не гарантированно верное решение, а лучшее из возможных.

Недостатки метода "ближайшего соседа":

– Данный метод не создает каких-либо моделей или правил, обобщающих предыдущий опыт, – в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на каком основании строятся ответы;

– Существует сложность выбора меры "близости" (метрики). От этой меры главным образом зависит объем множества записей, которые нужно хранить в памяти для достижения удовлетворительной классификации или прогноза. Также существует высокая зависимость результатов классификации от выбранной метрики;

– При использовании метода возникает необходимость полного перебора обучающей выборки при распознавании, следствие этого – вычислительная трудоемкость;

– Типичные задачи данного метода – это задачи небольшой размерности по количеству классов и переменных [13].

3.3. Нейронные сети. Нейронные сети (Neural Networks) – это модели биологических нейронных сетей мозга, в которых нейроны имитируются относительно простыми, часто однотипными, элементами (искусственными нейронами).

Нейронная сеть может быть представлена направленным графом с взвешенными связями, в котором искусственные нейроны являются вершинами, а синаптические связи – дугами.

Среди областей применения нейронных сетей – автоматизация процессов распознавания образов, прогнозирование, адаптивное управление, создание экспертных систем, организация ассоциативной памяти, обработка аналоговых и цифровых сигналов, синтез и идентификация электронных цепей и систем.

С помощью нейронных сетей решаются следующие задачи Data Mining:

– Классификация (обучение с учителем). Примеры задач классификации: распознавание текста, распознавание речи, идентификация личности;

– Прогнозирование. Для нейронной сети задача прогнозирования может быть поставлена таким образом: найти наилучшее приближение функции, заданной конечным набором входных значений (обучающих примеров). Например, нейронные сети позволяют решать задачу восстановления пропущенных значений;

– Кластеризация (обучение без учителя). Примером задачи кластеризации может быть задача сжатия информации путем уменьшения размерности данных. Задачи кластеризации решаются, например, самоорганизующимися картами Кохонена.

Недостатки нейронных сетей:

– Сложность может вызвать вопрос о количестве наблюдений в наборе данных. И хотя существуют некие правила, описывающие связь между необходимым количеством наблюдений и размером сети, их верность не доказана;

– Количество необходимых наблюдений зависит от сложности решаемой задачи. При увеличении количества признаков количество наблюдений возрастает нелинейно;

– Аналитик должен определить количество слоев в сети и количество нейронов в каждом слое. Алгоритма выбора оптимальной структуры до сих пор не существует;

– При обучении нейронных сетей часто возникает серьезная трудность, называемая проблемой переобучения (overfitting).

Переобучение, или чрезмерно близкая подгонка – излишне точное соответствие нейронной сети конкретному набору обучающих примеров, при котором сеть теряет способность к обобщению. Переобучение связано с тем, что выбор обучающего (тренировочного) множества является случайным. С первых шагов обучения происходит уменьшение ошибки. На последующих шагах с целью уменьшения ошибки (целевой функции) параметры подстраиваются под особенности обучающего множества. Однако при этом происходит "подстройка" не под общие закономерности ряда, а под особенности его части – обучающего подмножества. При этом точность прогноза уменьшается.

3.4. Метод опорных векторов. Метод опорных векторов (Support vector machines, SVM) был описан в работах В.Н. Вапника [2, 15]. SVM – это математический метод получения функции, решающей задачу классификации [16].

Идея метода возникла из геометрической интерпретации задачи классификации. Пусть два множества точек можно разделить плоскостью (в двумерном пространстве – прямой). Тогда таких плоскостей будет бесконечное множество (рисунок 2а). Выберем в качестве оптимальной такую плоскость, расстояния до которой ближайших точек обоих классов равны (рисунок 2б). Ближайшие точки-векторы называются опорными. Поиск оптимальной плоскости приводит к задаче квадратичного программирования при множестве линейных ограничений-неравенств. В 90-х гг. прошлого века метод SVM был усовершенствован: разработаны эффективные алгоритмы поиска оптимальной плоскости, найдены способы обобщения на нелинейные случаи и ситуации с числом классов, большим двух [17].

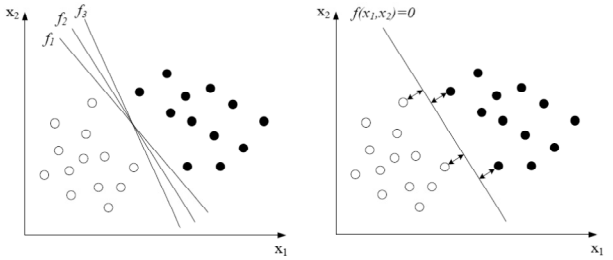

Рис. 2. Иллюстрация основной идеи SVM

Преимущества метода SVM:

– Метод опорных векторов, в отличие от нейронных сетей, устойчив к переобучению. Данный алгоритм может обучаться на выборке размером в гигабайты исходных данных, сильно коррелирующих между собой;

– работа с высокой размерностью входных векторов[18];

– конкурентоспособность по сравнению с методами, основанными на других алгоритмах.

Недостатки метода SVM:

– Метод опорных векторов неустойчив по отношению к шуму в исходных данных. Если обучающая выборка содержат шумовые выбросы, они будут существенным образом учтены при построении разделяющей гиперплоскости. Этого недостатка лишён метод релевантных векторов (relevance vector machine, RVM);

– До сих пор не разработаны общие методы построения спрямляющих пространств или ядер, наиболее подходящих для конкретной задачи. Построение адекватного ядра является искусством и, как правило, опирается на априорные знания о предметной области. На практике вполне разумные функции K(x, x′), выведенные из содержательных соображений, далеко не всегда оказываются положительно определёнными;

– В общем случае, когда линейная разделимость не гарантируется, приходится подбирать управляющий параметр алгоритма C [17].

Снижение размерности пространства признаков при помощи МГК:

Метод главных компонент – это итерационная процедура, в которой новые компоненты добавляются последовательно, одна за другой. Важно знать, когда остановить этот процесс, т.е. как определить правильное число главных компонент, A. Если это число слишком мало, то описание данных будет не полным. С другой стороны, избыточное число главных компонент приводит к

переоценке, т.е. к ситуации, когда моделируется шум, а не содержательная информация.

Результаты исследования, представленного в таблице 1, показали, что сокращение размерности не влияет критически на уровень верного распознавания нормальных и аномальных пакетов.

В итоговой реализации размерность с 41 признака была сокращена до 27 (на 34%).

Таблица 1. Оценка параметров эффективности работы алгоритма в пространстве признаков с полной и сниженной размерностью

Тип алгоритма	CR(%)
Без сокращения размерности	90,0
С сокращением размерности на 10%	89,7
С сокращением размерности на 15%	83,7
С сокращением размерности на 20%	82,4
С сокращением размерности на 30%	89,4
С сокращением размерности на 40%	76,5

INCremental Active Set method (INCAS). Алгоритм INCAS (INCremental Active Set method) позволяет уменьшить число вычислений при построении SVM. Для обучения SVM применяются алгоритмы, учитывающие специфические особенности SVM. Специфика заключается в том, что число опорных векторов h, как правило, невелико, $h \ll \ell$, и эти векторы находятся поблизости от границы классов. Именно эти особенности и позволяют ускорить поиск опорных объектов.

Эффективность:

– Оптимизационная задача зависит только от матриц Q и Qss. Следовательно, скалярные произведения надо вычислять только для пар "опорный-опорный" и "опорный-нарушитель";

– На каждой итерации IS к множеству добавляется только один объект. Значит, для пересчета обратной матрицы требуется меньше операций.

Преимущества:

– Метод позволяет решать задачи, где нет линейной разделимости;

– Алгоритм особенно эффективен, если число опорных векторов невелико;

– Данные могут поступать в режиме реального времени.

Недостатки:

– Алгоритм становится неэффективным, если число опорных векторов велико. В этом случае либо меняют ядро, либо саму постановку задачи.

Критерии эффективности работы алгоритма. Можно выделить следующие наиболее важные критерии оценки эффективности работы эвристических ССОВ (HNIDS):

– CR – количество корректно распознанных аномальных и нормальных пакетов; здесь также будет корректно предположить, что любые атаки обычно не входят в нормальный трафик сети и классифицируются как аномальные;

– FP (False Positive, ложная тревога) – количество нормальных пакетов принятых за аномальные;

– PPs (Packet per second, пакетооборот) – максимальное количество пакетов, которое система может обработать за 1 секунду на этапе тестирования;

– n (устойчивость системы) – процент отрицательных векторов в обучающей выборке, при котором система начинает работать нестабильно;

– FN (False Negative) – количество аномальных пакетов, принятых за нормальные [19].

3.5. Исследование существующих алгоритмов ССОВ. В условиях реальных современных сетей на применение ССОВ накладываются особые требования, связанные с высокими уровнями трафика (большими и сверхбольшими показателями пакетооборота в сети). Во-первых, скорость этапа тестирования имеет наивысший приоритет. Во-вторых, при проверке большого количества пакетов в секунду, любое ложное срабатывание вызывает появление сообщения в журналах аномалий. В таблице 2 указаны существующие ССОВ, которые удовлетворяют указанным выше условиям [20].

Таблица 2 Сравнение уровня детектирования и процента ложных срабатываний для различных алгоритмов

Метод	Основа	CR(%)	FP(%)
SPADE	временные закономерности	ок. 70	ок. 0.02
Геометрический подход Арнольда и Эскина	K ближайших соседей	89	10
fpMAFIA	адаптивная решётка	90,2	5,4
Кластеризация по Эскину	single linkage clustering	65,7	0,178
OTAD	одноклассовый SVM	59,1	3,1
PSO-SVM	одноклассовый SVM с предобработкой входящих данных и оптимизацией параметров	82,6	нет данных
Алгоритм с оптимизацией INCAS и МГК	одноклассовый SVM с предобработкой входящих данных и оптимизацией параметров	89,4	3

Если значение ложных срабатываний системы достаточно велико, то журналы системы очень быстро заполнятся ошибками распознавания и восприятие человеком настоящих аномалий в этом шуме будет сильно затруднено.

Другой важный момент заключается в том, что изначально не возможно разделить тренировочные данные на нормальные и аномальные (далее – положительные и отрицательные). То есть, тренировочная выборка может либо состоять целиком из данных, которые мы считаем положительными (или отрицательными), либо считается, что тренировочная выборка – смешанная.

4. Заключение. Современные методы обнаружения вторжений строятся на двух принципах: сигнатурный (формальный) и эвристический (обнаружение аномалий, основывающихся на моделях штатного функционирования наблюдаемой информационной системы).

Следует заметить, что существуют две крайности при использовании данной технологии:

– обнаружение аномального поведения, которое не является атакой, и отнесение его к классу атак (ошибка второго рода);

– пропуск атаки, которая не подпадает под определение аномального поведения (ошибка первого рода). Этот случай гораздо более опасен, чем ложное причисление аномального поведения к классу атак.

Поэтому при инсталляции и эксплуатации систем такой категории, обычные пользователи и специалисты сталкиваются с двумя довольно сложными задачами:

– построение профиля объекта — это трудно формализуемая и затратная по времени задача;

– определение граничных значений характеристик поведения субъекта для снижения вероятности появления одного из двух вышеназванных крайних случаев.

Обычно, системы обнаружения аномальной активности используют журналы регистрации и текущую деятельность пользователя в качестве источника данных для анализа. Достоинства систем обнаружения атак на основе технологии выявления аномального поведения можно оценить следующим образом:

– системы обнаружения аномалий способны выявлять новые типы атак, сигнатуры для которых еще не разработаны;

– они не нуждаются в обновлении сигнатур и правил обнаружения атак.

Недостатками систем на основе технологии обнаружения аномального поведения являются следующие:

– системы требуют длительного и качественного обучения;

– системы генерируют много ошибок второго рода;

– системы обычно слишком медленны в работе и требуют большого количества вычислительных ресурсов.

ССОВ на основе метода опорных векторов представляют собой перспективную тему для изучения, так как методы оптимизации параметров SVM применительно к специфике задачи на момент написания работы ещё малоизученны.

Мы рассматриваем существующие алгоритмы машинного обучения, такие как нейронные сети, метод K-ближайшего соседа, байесовские сети, а так же оцениваем их достоинства и недостатки;

Для реализации эвристических методов применим алгоритмы машинного обучения. Задача, поставленная перед алгоритмом, сводится к одноклассовой классификации с предварительным обучением на квазиположительной выборке. Поэтому для реализации был выбран алгоритм Support Vector Machines (SVM, машина опорных векторов) [2]

Использование данного метода применительно к текущей проблеме также представляет интерес, так как он является сравнительно малоизученным по сравнению с другими алгоритмами искусственного интеллекта.

Для минимизации размерности и повышения скорости обучения в качестве предобработки предлагаем метод главных компонент (МГК, англ. Principal component analysis, PCA) [3], позволяющий снизить время работы алгоритма без серьёзных потерь в точности обнаружения.

Для оптимизации параметров SVM предлагаем метод INCAS (INCremental Active Set method) [4] для данной задачи. В число преимуществ предложенного алгоритма обнаружения вторжений входит способность работать с любым аномальным трафиком, включая уязвимости нулевого дня. Аналогичный алгоритм может с успехом применяться на более высоких уровнях модели OSI (open system interconnection basic reference model), например, для защиты http-серверов от несанкционированного доступа.

Литература

1. *Котенко И.В., Воронцов В.В., Чечулин А.А., Уланов А.В.* Проактивные механизмы защиты от сетевых червей: подход, реализация и результаты экспериментов // Информационные технологии. 2009. №. 1. С. 37–42.
2. *Cortes C., Vapnik V.* Support vector networks // Machine Learning. 1995. vol. 20. 273–297.
3. *Jolliffe I.T.* Principal components analysis // New York: Springer-Verlag. 1986. 487 p.
4. *Fine S., Scheinberg K.* INCAS: An incremental active set method for SVM // Tech. rep.: 2002.
5. *Handley M., Kreibich C., Paxson V.* Network Intrusion Detection: Evasion, Traffic Normalization // Proc. 10th USENIX Security Symposium. 2001. pp. 115-131.

6. *Ferguson P., Senie D.* Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing // Internet Engineering Task Force RFC 2267. 1998. URL: http://www.ietf.org/rfc/rfc2267.txt.
7. *Котенко И.В., Саенко И.Б., Полубелова О.В., Чечулин А.А.* Применение технологии управления информацией и событиями безопасности для защиты информации в критически важных инфраструктурах // Труды СПИИРАН. Вып. 1(20). 2012. С. 27–56.
8. *Kotenko I.V., Chechulin A.A.* A Cyber Attack Modeling and Impact Assessment Framework // Proceedings of 5th International Conference on Cyber Conflict 2013 (CyCon 2013). IEEE and NATO COE Publications. 2013. pp. 119–142.
9. *Чечулин А.А., Котенко И.В.* Комбинирование механизмов защиты от сканирования в компьютерных сетях // Информационно-управляющие системы. 2010. № 12. С.21–27.
10. *Тулупьев А.Л.* Задача локального автоматического обучения в алгебраических байесовских сетях: логико-вероятностный подход // Труды СПИИРАН. 2008. Вып. 7. С. 10–25.
11. *Городецкий В.И.* Алгебраические байесовские сети — новая парадигма экспертно-вычислительных систем // Юбилейный сборник трудов институтов Отделения информатики, вычислительной техники и автоматизации РАН. М.: РАН. 1993. Т. 2. С. 120–141.
12. *Zaidi N.A., Squire D.M., Suter D.* A gradient-based metric learning algorithm for k-nn classifiers // Proceedings of the Australasian Joint Conference on Artificial Intelligence. 2011. vol. 6464. pp. 194-203.
13. *Weinberger K.Q., Blitzer J., Saul L.K.* Distance metric learning for large margin nearest neighbor classification. // Proceedings of Neural Inforamtion and Processing Systems. 2006.
14. *Котенко И.В., Дойникова Е.В., Чечулин А.А.* Общее перечисление и классификация шаблонов атак (CAPEC): описание и примеры применения // Защита информации. Инсайд. 2012. № 4. С. 54-66.
15. *Zhao C., Wang H. G. Cai.* Study on a SVM-based Data Fusion Method, Proceedings of IEEE Conference on Robotics // Automation and Mechatronics. 2004 pp. 413–415.
16. *Vapnik V.N*. Statistical Learning Theory // Wiley. 1998. 768 p.
17. *Bartlett P., Shawe-Taylor J.* Generalization performance of support vector machines // Advances in Kernel Methods – Support Vector Learning. MIT Press. 1999. pp. 43–54.
18. *Collobert R., Bengio Y., Bengio S.* A parallel mixture of SVMs for very large-scale problems // Neural Information Processing Systems. Advances in Neural Information Processing Systems. MIT Press. 2002. vol. 14. pp. 633-640.
19. *Большев А.К.* Алгоритмы преобразования и классификации трафика для обнаружения вторжений в компьютерные сети: автореф. // Санкт-Петербург: б.н. 2011. 19 с.
20. *Wang J., Hong X., Ren R.R., Li T.* A Real-time Intrusion Detection System Based on PSO-SVM // Proceedings of the 2009 International Workshop on Information Security and Application (IWISA 2009). Qingdao. China. 2009. pp. 319-321.
21. *Gorodetsky V.I., Kotenko I.V., Karsaev O.I.* Multi-agent Technologies for Computer Network Security: Attack Simulation, Intrusion Detection and Intrusion Detection Learning // International Journal of Computer Systems Science and Engineering. 2003. vol. 18. no. 4. pp. 191–200.
22. *Bitter C., North J., Elizondo D. A., Watson T.* An Introduction to the Use of Neural Networks for Network Intrusion Detection // Computational Intelligence for Privacy and Security Studies in Computational Intelligence. 2012. vol. 394. pp. 5–24.
23. *Jing X.* IDS Method Based on Improved SVM Algorithm Under Unbalanced Data Sets // Proceedings of the 2012 International Conference on Cybernetics and InformaticsLecture Notes in Electrical Engineering. 2014. vol. 163. pp. 413–420.

References

1. Kotenko I.V., Voroncov V.V., Chechulin A.A., Ulanov A.V. [Proactive security mechanisms against network worms: approach, implementation and results of the experiments] *Informacionnye tehnologii – Information Technology*. 2009. vol. 1. pp. 37–42. (In Russ.).
2. Cortes C., Vapnik V. Support vector networks. Machine Learning. 1995. vol. 20. 273–297.
3. Jolliffe I.T. Principal components analysis. New York: Springer-Verlag. 1986. 487 p.
4. Fine S., Scheinberg K. INCAS: An incremental active set method for SVM. Tech. rep.: 2002.
5. Handley M., Kreibich C., Paxson V. Network Intrusion Detection: Evasion, Traffic Normalization. Proc. 10th USENIX Security Symposium. 2001. pp. 115-131.
6. Ferguson P., Senie D. Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing. Internet Engineering Task Force RFC 2267. 1998. URL: http://www.ietf.org/rfc/rfc2267.txt.
7. Kotenko I.V., Saenko I.B., Polubelova O.V., Chechulin A.A. [The application of information management technologies and security events to protect the information in critical infrastructures]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2012. vol. 1(20). pp. 27–56. (In Russ.).
8. Kotenko I.V., Chechulin A.A. A Cyber Attack Modeling and Impact Assessment Framework. Proceedings of 5th International Conference on Cyber Conflict 2013 (CyCon 2013). IEEE and NATO COE Publications. 2013. pp. 119–142.
9. Chechulin A.A., Kotenko I.V. [The combination of mechanisms of protection against scanning computer networks] *Informacionno-upravljajushhie sistemy – Information and Control Systems*. 2010. vol. 12. pp. 21–27. (In Russ.).
10. Tulup'ev A.L. [The task of the local auto-learning algebraic Bayesian network: logical-probabilistic approach]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2008. vol. 7. pp. 10–25. (In Russ.).
11. Gorodeckij V.I. [Algebraic Bayesian network - a new paradigm of computing systems expert] *Jubilejnyj sbornik trudov institutov Otdelenija informatiki, vychislitel'noj tehniki i avtomatizacii RAN – Anniversary Proceedings of the Institute of the Department of Informatics, Computer Science and Automation RAS*. M.: RAN. 1993. vol. 2. pp. 120–141. (In Russ.).
12. Zaidi N.A., Squire D.M., Suter D. A gradient-based metric learning algorithm for k-nn classifiers. Proceedings of the Australasian Joint Conference on Artificial Intelligence. 2011. vol. 6464. pp. 194–203.
13. *Weinberger K.Q., Blitzer J., Saul L.K.* Distance metric learning for large margin nearest neighbor classification. Proceedings of Neural Information and Processing Systems. 2006.
14. Kotenko I.V., Dojnikova E.V., Chechulin A.A. [Total enumeration and classification of attack patterns (TSAPETS): description and application examples] *Zashhita informacii. Insajd – Protection of information. Inside*. 2012. vol. 4. pp. 54–66. (In Russ.).
15. Zhao C., Wang H. G. Cai. Study on a SVM-based Data Fusion Method. Proceedings of IEEE Conference on Robotics. Automation and Mechatronics. 2004 pp. 413–415.
16. Vapnik V.N. Statistical Learning Theory. Wiley. 1998. 768 p.
17. Bartlett P., Shawe-Taylor J. Generalization performance of support vector machines. Advances in Kernel Methods – Support Vector Learning. MIT Press. 1999. pp. 43–54.
18. Collobert R., Bengio Y., Bengio S. A parallel mixture of SVMs for very large-scale problems. Neural Information Processing Systems. Advances in Neural Information Processing Systems. MIT Press. 2002. vol. 14. pp. 633–640
19. Bol'shev A.K. *Algoritmy preobrazovanija i klassifikacii trafika dlja obnaruzhenija vtorzhenij v komp'juternye seti* [Conversion algorithms and classification of traffic for intrusion detection in computer networks]. St. Petersburg. 2011. 19p. (In Russ.).

20. Wang J., Hong X., Ren R.R., Li T. A Real-time Intrusion Detection System Based on PSO-SVM. Proceedings of the 2009 International Workshop on Information Security and Application (IWISA 2009). Qingdao. China. 2009. pp. 319-321.
21. Gorodetsky V.I., Kotenko I.V., Karsaev O.I. Multi-agent Technologies for Computer Network Security: Attack Simulation, Intrusion Detection and Intrusion Detection Learning. International Journal of Computer Systems Science and Engineering. 2003. vol. 18. no. 4. pp. 191–200.
22. Bitter C., North J., Elizondo D. A., Watson T. An Introduction to the Use of Neural Networks for Network Intrusion Detection. Computational Intelligence for Privacy and Security Studies in Computational Intelligence. 2012. vol. 394. pp. 5–24.
23. Jing X. IDS Method Based on Improved SVM Algorithm Under Unbalanced Data Sets. Proceedings of the 2012 International Conference on Cybernetics and InformaticsLecture Notes in Electrical Engineering. 2014. vol. 163. pp. 413–420.

Носков Антон Николаевич — старший преподаватель, кафедра компьютерной безопасности математического факультета Ярославского государственного университета им. П.Г. Демидова. Область научных интересов: сетевая безопасность, сети, протоколы маршрутизации, Cisco. Число научных публикаций — 4. lantoni@uniyar.ac.ru; 150000, г. Ярославль, ул. Советская-14. каб 225; р.т. +79106641047.

Noskov Anton Nikolaevich — senior teacher, Department of Computer Security of The Mathematics Faculty of P.G.Demidov Yaroslavl State University. Research interests: networks, routing protocols, network security. The number of publications — 4. lantoni@uniyar.ac.ru; 1500006 Yaroslavl, Sovetskaya-14, 225; office phone +79106641047.

Чечулин Андрей Алексеевич — к-т техн. наук, старший научный сотрудник, лаборатория проблем компьютерной безопасности Федерального государственного бюджетного учреждения науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук. Область научных интересов: безопасность компьютерных сетей, обнаружение вторжений, анализ сетевого трафика, анализ уязвимостей. Число научных публикаций — 110. andreych@bk.ru, http://comsec.spb.ru/ru/staff/chechulin; 199178, г. Санкт-Петербург, 14-я линия В.О., д.39, комната 215; р.т. +78123287181.

Chechulin Andrey Alexeevich — Ph.D., senior researcher, laboratory of Computer Security Problems of the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Science. Research interests: computer network security, intrusion detection, analysis of the network traffic, vulnerability analysis. The number of publications — 110. andreych@bk.ru, http://comsec.spb.ru/ru/staff/chechulin; 199178, Saint-Petersburg, liniya 14-ya, 39, room 215; office phone +78123287181.

Тарасова Дарья Алексеевна — студент физический факультет Ярославского государственного университета им. П.Г. Демидова. Область научных интересов: сети, протоколы маршрутизации, методы коммутации. Число научных публикаций — 3. lantoni@mail.ru; 150000, г. Ярославль, Советская-14.; р.т. +7 4852 797725.

Tarasova Daria Alekseevna — student Physics Faculty of P.G.Demidov Yaroslavl State University. Research interests: networks, routing, switching. The number of publications — 3. lantoni@mail.ru; 150000, Yaroslavl, Sovetskaya-14; office phone +7 4852 797725.

Поддержка исследований. Работа выполнена при финансовой поддержке РФФИ (проекты №13-01-00843, 13-07-13159, 14-07-00697, 14-07-00417 и 14-37-50735), программой фундаментальных исследований ОНИТ РАН (контракт №2.2) и проектом ENGENSEC программы Европейского Сообщества TEMPUS.

Acknowledgements. This research is supported by RFBR (grants #13-01-00843, 13-07-13159, 14-07-00697, 14-07-00417 и 14-37-50735), by the ONIT RAS (project #2.2) as well as by the project ENGENSEC of the European Community program TEMPUS.

РЕФЕРАТ

Носков А.Н., Чечулин А.А., Тарасова Д.А. **Исследование эвристических подходов к обнаружению атак на телекоммуникационные сети на базе методов интеллектуального анализа данных.**

Одна из научных проблем связанных с безопасностью систем, это разработка методологических основ обеспечения безопасности сетей и обнаружения нежелательного трафика. Современные методы обнаружения вторжений строятся на двух принципах: сигнатурный (формальный) и эвристический (обнаружение аномалий, основывающихся на моделях штатного функционирования наблюдаемой информационной системы). В статье дан анализ существующих методов машинного обучения сетевых угроз, применимых для решения задачи обнаружения сетевых атак. Одним из перспективных методов решения такой задачи является метод опорных векторов. Для минимизации размерности и повышения скорости обучения в качестве предобработки предложен метод главных компонент, позволяющий снизить время работы алгоритма без серьёзных потерь в точности обнаружения. Для оптимизации параметров SVM для данной задачи предложен метод INCAS (INCremental Active Set method).

SUMMARY

Noskov A.N., Chechulin A.A., Tarasova D.A. **Investigation of Heuristic Approach to Attacks on the Telecommunications Network Detection based on Data Mining Techniques.**

One of the scientific challenges which related to network security is the development of the methodological foundations of network security and detection of malicious and unwanted traffic. Current methods of intrusion detection systems are based on two principles: signature (formal) and heuristic (anomaly detection). The paper presents the analysis of the existing machine learning methods that can be applied to enhance the quality of intrusion detection systems. One of the perspective methods for solving this problem is a support vector machine. To minimize the dimension and improve the speed of learning as preprocessing a principal component analysis was proposed. This approach allows to reduce the time of the algorithm without serious losses of accuracy detection. For optimizing the parameters of SVM a method INCAS (INCremental Active Set method) is proposed.