УДК 004.056.57

ИССЛЕДОВАНИЕ МЕТОДОВ И АЛГОРИТМОВ ОБНАРУЖЕНИЯ СЕТЕВЫХ АНОМАЛИЙ

Богиня И. Г., Масальская Е. А.
ГВУЗ «Национальный горный университет», http://bit.nmu.org.ua/, E-mail: big94@ua.fm

В данной статье рассматриваются основные методы и алгоритмы обнаружения сетевых аномалий и атак, проводится сравнительный анализ данных алгоритмов, а также рассматривается применимость этих алгоритмов в качестве базы для создания системы обнаружения сетевых аномалий.

Ключевые слова – сетевые аномалии; системы обнаружения аномалий.

ВСТУПЛЕНИЕ

Одной из актуальных задач в сфере информационной безопасности является создание системы обнаружения нестандартной сетевой активности. Реализацию данной задачи осложняет тот факт, что почти каждый день появляются новые виды сетевых атак и инструментов воздействия на объекты сетей, что приводит к ошибкам в работе систем обнаружения вторжений (СОВ) и обнаружения аномалий (СОА). При анализе состояния сети, для минимизации ошибок, генерируемых СОВ и СОА, привлекают экспертов в области ИБ. Однако, присутствие человека в системе в качестве звена, принимающего какие-либо решения, может приводить к повышению числа невзвешенных и импульсивных решений в виду особенностей человеческого поведения.

Для снижения влияния человеческого фактора на процессы почти любой производственной сферы внедряют автоматизацию данных процессов. В сфере информационной безопасности важными процессами, влияющими на состояние информации и рассматриваемой информационной системы, являются процессы мониторинга состояния всех информационных ресурсов, а также процессы реакции на обнаруженные аномалии. В данном случае, под аномалиями подразумеваются любые отклонения состояний объектов системы от эталонных для таковых.

ОСНОВНАЯ ЧАСТЬ

Обнаружение аномалий – динамический метод работы антивирусов, хостовых и сетевых систем обнаружения вторжений. Программное обеспечение, использующее этот метод, наблюдает определённые действия (работу программы/процесса, параметры сетевого трафика, работу пользователя), следя за возможными необычными и подозрительными событиями или тенденциями.

Одним из важных аспектов необходимости проведения мониторинга работы ИС является то, что наличие аномалии может указывать на проводимую в настоящем времени атаку на информационные ресурсы данной ИС. Рынок программных и программно-аппаратных продуктов, задачей которых являются мониторинг состояния ИС и осуществление реакций на возникновение аномалий, насчитывает немало позиций таких, как: Catchi, Zabbix, Nagios, Ganglia, Snort и другие [1].

В случаях, когда в ИС возникает аномалия и необходимо предпринять ответные действия, которые не могут быть просчитаны ни одной из имеющихся СОВ, привлекают экспертов в области ИБ. Но, если речь идет о реализации угроз для критической информации, время для ответных действий по защите сокращается практически до нуля, что накладывает существенные временные ограничения на процессы привлечения экспертов. Из-за этого возникает необходимость создания системы, способной реагировать на возникающие в ИС аномалии в режиме реального времени. Некоторые из вышеприведенных продуктов справляются с задачей не только мониторинга, но и обнаружения и предотвращения вторжений и атак, но качество обнаружения атак напрямую зависит от вида проводимой атаки. Некоторые СОВ работают по сигнатурному принципу, что позволяет хорошо обнаруживать множество распространенных видов атак, для которых собрано много информации, описывающей данные атаки. Но недостатком таких систем является их неспособность дать адекватный ответ на атаку, для который в базе знаний СОВ нет никакой информации. В таких случаях используют СОВ, основанные на эвристических алгоритмах. Однако, такие системы также не могут работать в полном диапазоне возможных видов атак из-за самого определения эвристических методов поиска решений: данные алгоритмы не имеют доказанной правильной логики для всех возможных вариантов решаемых задач.

Одним из возможных вариантов построения данной системы является моделирование поведения эксперта в сфере ИБ в ситуации возникновения аномалии. В подавляющем большинстве случаев, логика мышления человека, который должен рассуждать рационально и взвешенно принимать решения, возможно описать математическими моделями такими, как деревья принятия решений [2].

В широком смысле, деревья принятия решений – средство поддержки принятия решений, использующееся в статистике и анализе данных для прогнозных моделей. Структура дерева представляет собой «листья» и «ветки». На ребрах («ветках») дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах – атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, необходимо спуститься по дереву до листа и выдать соответствующее значение. Цель состоит в том, чтобы

создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе.

Существуют различные реализации данной модели, такие как: алгоритмы CART, C4.5, CHAID, CN2, NewId, ITrule и т.д. Наиболее часто из них применяются алгоритмы CART и C4.5 [3].

Алгоритм CART принципиально отличается от некоторых других алгоритмов конструирования деревьев решений механизмом отсечения ветвей. В рассматриваемом алгоритме отсечение – это некий компромисс между получением дерева «подходящего размера» и получением наиболее точной оценки классификации. Также для применения алгоритма CART нет необходимости заранее выбирать переменные, которые будут участвовать в анализе: переменные отбираются непосредственно во время проведения анализа на основании значения индекса Джини (показателя неравномерности распределения некоторой величины на заданном интервале).

Однако, данный алгоритм не лишен недостатков. В случае, когда необходимо построить дерево со сложной структурой, лучше использовать другие алгоритмы, т.к. CART может не идентифицировать правильную структуру данных.

Алгоритм C4.5 является усовершенствованной версией алгоритма ID3. В частности, в новую версию были добавлены отсечение ветвей, возможность работы с числовыми атрибутами, а также возможность построения дерева из неполной обучающей выборки, в которой отсутствуют значения некоторых атрибутов.

В обучающей выборке количество примеров должно быть значительно больше количества классов, к тому же каждый пример должен быть заранее ассоциирован со своим классом. По этой причине C4.5 является вариантом машинного обучения с учителем.

Одним из алгоритмов, адаптированных для максимально гетерогенных входящих данных, является алгоритм Random forest – алгоритм машинного обучения, предложенный Лео Брейманом и Адель Катлер, заключающийся в использовании комитета (ансамбля) решающих деревьев. Алгоритм применяется для задач классификации, регрессии и кластеризации [4].

Классификация объектов проводится путём голосования: каждое дерево комитета относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев.

Оптимальное число деревьев подбирается таким образом, чтобы минимизировать ошибку классификатора на тестовой выборке. В случае её отсутствия, минимизируется оценка ошибки «out-of-bag»: доля примеров обучающей выборки, неправильно классифицируемых комитетом, если не учитывать голоса деревьев на примерах, входящих в их собственную обучающую подвыборку.

Главными достоинствами данного алгоритма являются:

- способность эффективно обрабатывать данные с большим числом признаков и классов;
- нечувствительность к масштабированию (и вообще к любым монотонным преобразованиям) значений признаков;
- одинаково хорошо обрабатываются как непрерывные, так и дискретные признаки. Существуют методы построения деревьев по данным с пропущенными значениями признаков;
- существуют методы оценивания значимости отдельных признаков в модели;
- внутренняя оценка способности модели к обобщению (тест «out-of-bag»);
- высокая параллелизуемость вычислений и масштабируемость.

Учитывая вышеописанные особенности алгоритма Random forest, а также его способность воспринимать широкий диапазон входных данных, данный алгоритм является наиболее подходящим для реализации в качестве основы системы поиска и ответа на обнаруженные аномалии в ИС.

Применять данный алгоритм возможно на нескольких этапах формирования реакции на возникшую аномалию. Сначала необходима кластеризация данных, полученных от модуля мониторинга разрабатываемой СОВ. После этого возможно непосредственное создание набора ответных действий, которое также осуществляется при помощи данных, представляемых на выходе моделью «случайного леса».

Необходимо отметить, что адекватность создаваемых реакций будет напрямую зависеть от объема и достоверности данных, полученных от модуля мониторинга СОВ.

ЗАКЛЮЧЕНИЕ

В качестве базы для создания описанной системы реагирования на аномалии рекомендуется выбирать математические модели такие, как модель комитета деревьев решений Random forest, т.к. в сравнении с вышеперечисленными моделями данный алгоритм наименее требователен к типизации входных данных (что свойственно специфике работы системы обнаружения аномалий в случае обнаружения нового типа аномалии), а также является особо эффективным при обработке данных с большим количеством признаков, что, соответственно, существенно повышает точность выходных данных.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Шелухин О.И., Сакалема Д.Ж., Филинова А.С. Обнаружение вторжений в компьютерные сети (сетевые аномалии). 2-е изд. М.: КМК, 2016. – 223 с.

2. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям: учебное пособие. 2-е изд. СПб: Питер, 2013. – 704 с.

3. Толстова Ю.Н. Анализ социологических данных. М.: Научный мир, 2000. - 352 с.

4. Проталинский О.М. Применение методов искусственного интеллекта при автоматизации технологических процессов. АГТУ, 2004. – 183 с.