МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ, КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ В ТЕОРИИ УПРАВЛЕНИЯ СЛОЖНЫХ ПРОЦЕССОВ

ПРИМЕНЕНИЕ СТАТИСТИЧЕСКОГО АНАЛИЗА СЕТЕВОГО ТРАФИКА С ИСПОЛЬЗОВАНИЕМ БАЙЕСОВСКОГО КЛАССИФИКАТОРА В ИНТЕРЕСАХ БЕЗОПАСНОСТИ ВЫЧИСЛИТЕЛЬНЫХ СЕТЕЙ

И.Г. Малыгин, доктор технических наук, профессор.
Санкт-Петербургский университет ГПС МЧС России.
Д.В. Козьмовский, кандидат технических наук.
Институт проблем транспорта им. Н.С. Соломенко
Российской академии наук

Рассмотрены методы обеспечения безопасности сетевых объектов управления и наиболее распространенные методы, которые реализованы в современных системах защиты. Описан метод защиты вычислительных сетей, основанный на методе статистического анализа трафика вычислительных сетей.

Ключевые слова: вычислительные сети, сетевой трафик, безопасность вычислительных сетей, статистический анализ трафика, байесовский классификатор

STATISTICAL ANALYSIS OF THE NETWORK TRAFFIC USING THE BAYESIAN CLASSIFIER IN THE INTERESTS OF SECURITY OF COMPUTER NETWORKS

I.G. Malygin. Saint-Petersburg university of State fire service of EMERCOM of Russia.
D.V. Kozmovsky. Institute of transportation problems of name N.S. Solomenko of the Russian academy of sciences

Considered the methods of ensuring the security of the network objects management. Described the most common methods implemented in modern protection systems. The described method of protection of computer networks based on a method of statistical analysis of the traffic area networks.

Keywords: computer network, network traffic, security of computer networks, statistical traffic analysis, Bayesian classifier

В настоящее время существует глубокая интеграция сетевых технологий в повседневную деятельность объектов управления всех уровней. Это может быть локальная сеть местного назначения, объединяющая несколько компьютеров небольшого узла связи или состоящая из нескольких сегментов сеть вычислительного центра с возможностью доступа в глобальную сеть Интернет.

Организация локальной сети сопряжена с необходимыми настройками политики безопасности пользователей сети. Использование сетевого экрана для защиты от сетевых атак, разграничение прав доступа пользователей к компьютерам, контроль подключений к сети извне – немногие трудности, которые встают перед администратором локальной сети.

Для решения этих задач проводится ряд мероприятий, например:

– настройка сетевых протоколов в соответствии с поставленными задачами;

– настройка политик прав доступа пользователей, а также средств аутентификации пользователей на каждом компьютере сети;

– анализ и контроль структуры локальной сети;

– организация доступа пользователей к компьютерам административными мерами.

Часть этих задач успешно решается достаточно известными аппаратно-программными средствами защиты информации, которые используются для блокировки клавиатуры, блокировки загрузки с внешних носителей, аутентификации и идентификации пользователей на начальном этапе загрузки компьютеров. Подобные средства обеспечивают защиту информации, хранящейся на жестком диске и на удаленных сетевых ресурсах, от несанкционированного доступа со стороны злоумышленников путем раздельного и гибко настраиваемого разграничения доступа к файлам, каталогам, принтерам, устройствам ввода/вывода и системным компонентам операционной системы. Обеспечивается защита системных компонентов операционной системы от случайной или умышленной модификации, удаления, поражения вирусами.

Внутренний нарушитель представляет собой легитимного пользователя вычислительной сети, который обладает определенными правами на доступ к информационным ресурсам. Вследствие умышленных или ошибочных действий внутренний нарушитель может принести ущерб, зачастую больший, чем внешний злоумышленник. На внутренние угрозы довольно часто не обращают внимания, полагая, что защититься полностью все равно невозможно, а лоскутная безопасность не принесет должного эффекта. Защита информации, циркулирующей в вычислительной сети, является важной задачей, и, учитывая, что более 90 % информации ныне находится в электронном виде, физические средства и методы защиты информации утратили свою былую эффективность.

Таким образом, даже при качественной организации и правильной настройке сети существует необходимость контроля деятельности пользователей. Поэтому необходимо проводить контроль сетевых соединений, учет и анализ сетевого трафика.

Системы, предназначенные для защиты вычислительной сети от внутреннего нарушителя, представляются следующими классами [1]:

- системы контроля доступа к локальным портам и сетевым интерфейсам;
- системы контроля доступа на основе политик безопасности;
- системы шифрования;
- системы DLP-класса – контентная фильтрация трафика;
- системы IRM-класса – гибрид системы контроля доступа и криптосистем;
- системы IDS-класса – сигнатурный анализ трафика.

Системы, представленные в трех первых классах, не имеют качественной защиты информации от легитимного пользователя вычислительной сети, главным образом, обеспечивая борьбу с внешним злоумышленником. Системы, функционирующие на основе контентного анализа, на сегодняшний день находятся на начальном этапе развития и не могут обеспечить высокого уровня защиты информации в отдельном исполнении. В свою очередь сигнатурный анализ имеет высокую степень распространения, часто применяется, особенно в антивирусных системах. Однако в рамках защиты информации, циркулирующей внутри вычислительной сети, системы сигнатурного анализа обладают рядом ограничений: необходимостью маркирования документов специальными метками, обновления сигнатур и увеличения ресурсоемкости данных систем в процессе ее использования. Кроме того, внедрение подобных систем является дорогостоящим и трудоемким мероприятием,

требующим продолжительного времени настройки.

Существующие системы анализа трафика позволяют проводить мониторинг трафика и сетевых соединений. В состав систем входят инструменты расшифровки пакетов сетевого обмена, возможность настройки фильтров по большому количеству критериев. Ведется учет трафика по узлам, вывод статистики входящего/исходящего трафика. Большинство систем предлагают классификацию трафика, основанную на нескольких методах:

– анализ открытых портов соединений основывается на том, что приложения работают по умолчанию на известных портах;

– анализ полезной нагрузки пакетов сетевого трафика заключается в обнаружении определенных сигнатур, специфичных для сетевых приложений, в полезной нагрузке пакетов.

К недостаткам первого метода можно отнести то, что большинство приложений позволяют изменять номера портов по умолчанию на любые. Многие современные приложения предпочитают использовать случайные номера портов. Также существует тенденция использования номеров портов известных приложений.

Недостатками второго метода являются необходимость ведения и обновления сигнатур, используемых сетевыми приложениями для более точного анализа трафика; при шифровании трафика анализ сильно затруднен; требуется анализ всего сетевого трафика, что может создать большие нагрузки на оборудование и вызвать ошибки в работе сети.

В целях обеспечения защищенности компьютерных сетей созданы системы, которые классифицируют сетевую активность различных программ. Однако большинство современных систем сетевой безопасности не имеют возможности самообучения и оперируют только заложенными в них вручную правилами. Применение методов интеллектуального анализа данных позволяет ввести в системы защиты свойство самообучения и обеспечивает обнаружение нежелательного сетевого трафика, основываясь на закономерностях, выявленных статистическим путем.

В результате существует необходимость разработки метода анализа трафика, который не использует привязку сетевых приложений к портам и не проводит анализ содержимого пакетов сетевого обмена, при этом показывая высокий показатель эффективности правильной классификации трафика.

При описании метода используется понятие «сеанса», то есть периода сетевого обмена пакетами между двумя узлами в течение непрерывного времени. Весь трафик будет поделен на сеансы, каждый из которых классифицируется отдельно, то есть вычисляется вероятность принадлежности сеанса тому или иному протоколу. Таким образом, трафик – это совокупность сеансов.

Поэтому предлагается метод классификации сеансов по видам деятельности, который позволит автоматизировать процесс анализа. Метод классификации заключается в выявлении статистических характеристик сеанса – показателей, по которым существует возможность классификации. Статистические данные сеансов берутся из заголовков пакетов сетевого обмена.

Экспериментальным путем был сформирован перечень статистических характеристик сеансов, которые можно рассматривать как независимые и, при этом, успешно применять для классификации трафика. К числу указанных характеристик относятся следующие [2]:

– число внешних портов;

– число внутренних портов;

– доля исходящего трафика;

– средний размер пакета.

Статистические параметры сеанса, которые выбраны для классификации трафика по видам сетевой деятельности, являются непрерывными числовыми величинами. Сложность использования метода Байеса заключается в том, что данный метод может работать либо с дискретными значениями, либо используя плотности распределения непрерывных величин. В случае использования первого варианта необходимо описать

переход от фактических значений статистических параметров сеанса к вектору значений дискретных параметров классификации $X = (x_1, x_2, \ldots, x_m)$.

Рассмотрим ситуацию, когда уже имеется некоторая накопленная статистика значений статистических параметров сеансов по различным видам сетевой деятельности. Значения статистических параметров можно представить в виде распределения каждого из параметров сеанса на отрезке возможных значений. В качестве примера приведем распределение параметра «Количество внутренних портов» для вида деятельности Web-серфинг (рис.):

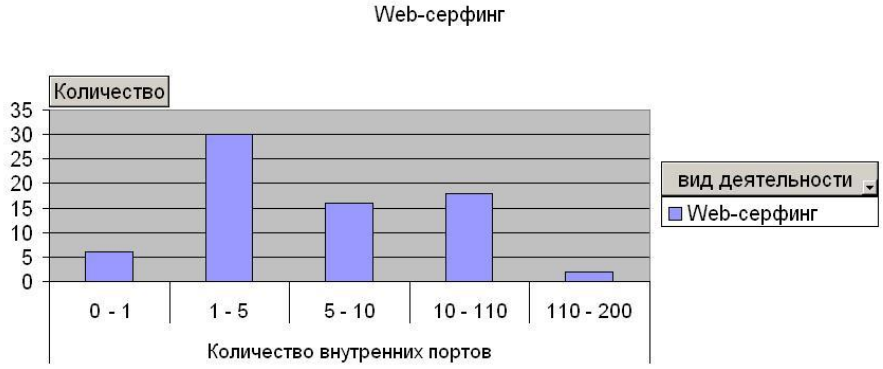

Рис. **Распределение значений статистического параметра «Количество внутренних портов» для вида деятельности Web-серфинг**

На приведенной диаграмме конкретные значения статистического параметра разделяются на интервалы значений. Интервалы статистических параметров сеансов формируются вручную в течение наполнения обучающей выборки, которое заключается в сборе статистической информации, то есть значений статистических параметров, по всем видам деятельности, для которых поставлена задача классификации. Количество интервалов и их размерность определяется распределением статистических данных на всем промежутке возможных значений. Процесс формирования интервалов происходит от общего к частному таким образом, чтобы, по возможности, получить распределения значений параметров разных видов деятельности по интервалам, при этом, не допуская излишней дискретности, которая может усложнить задачу классификации.

Переход от фактических значений статистических параметров сеанса к значениям интервалов объясняется необходимостью дискретности данных, используемых в «Наивном байесовском классификаторе» (НБК). В результате вместо фактических значений параметров в качестве параметров классификации используются интервалы промежутков возможных значений статистических параметров. Для функционирования данного метода необходимо наличие обучающей выборки по протоколам, которые будут подвергаться классификации. При наличии большой выборки по каждому протоколу, вероятность точной классификации достаточно высока.

В качестве математического аппарата предлагается использование метода классификации объектов, известного как НБК [3], для вероятностной оценки соответствия сетевых сеансов тем или иным видам сетевой деятельности. Указанный метод классификации широко используется в системах для определения нежелательных почтовых сообщений (анти-спам системах). После некоторой модификации возможно использование НБК для классификации практически любых видов сетевых сеансов, при условии, что в распоряжении эксперта имеется достаточная по объему обучающая выборка с образцами подобных сеансов. В основе метода лежит формула условной вероятности:

$$P(H_i \mid X) = \frac{P(X \mid H_i)P(H_i)}{\sum_{k=1}^{n} P(X \mid H_k)P(H_k)},$$

где $P(H_i|X)$ – вероятность истинности гипотезы H_i при заданной причине X; $P(H_i)$ – априорная вероятность гипотезы H_i; $P(X|H_i)$ – вероятность присутствия причины X, если истинна гипотеза H_i; n – число возможных гипотез.

Если причину можно представить в виде вектора: $X=(x_1, x_2, \ldots, x_m)$, каждый компонент которого имеет условную вероятность относительно гипотезы H_i $P(x_j|H_i)$, то для вычисления условных вероятностей $P(X|H_i)$ используется «наивное» предположение об условной независимости компонентов вектора X [4]. В этом случае условная вероятность вычисляется по формуле:

$$P(X \mid H_i) = \prod_{j=1}^{m} P(x_j \mid H_i).$$

В случае классификации сетевых сеансов в качестве гипотез H_i выступают предположения о том, что классифицируемый сеанс соответствует i-му виду сетевой деятельности. То есть вероятность соответствия сеанса с набором параметров классификации X виду деятельности H_i равна произведению вероятностей соответствия каждого вычисленного параметра классификации виду деятельности H_i.

Применение выбранного метода классификации на основе НБК предполагает использование статистической информации. Объем статистической информации, который содержит в себе сеансы точно определенных видов деятельности, называется обучающей выборкой. А сеансы, которые содержит обучающая выборка, называются эталонными. Для каждого эталонного сеанса произведен расчет статистических значений параметров и произведено распределение данных значений по интервалам. Для каждого нового сеанса, который подвергается классификации, вычисляются значения параметров классификации и определяются диапазоны, в которые попали рассчитанные значения. Данные номера диапазонов являются параметрами классификации. Расчет соответствия параметра классификации x_j виду деятельности H_i производится по формуле:

$$P(x_j|H_i) = m/M ,$$

где M – общее количество значений данного параметра x_j в обучающей выборке, накопленное эталонными сеансами вида деятельности H_i; m – количество значений данного параметра x_j, накопленное эталонными сеансами вида деятельности H_i, в диапазоне, в который попадает вычисленное значение параметра классифицируемого сеанса.

Результаты применения предложенного метода представлены в таблице [5].

Таблица. **Результаты проведения классификации**

Результаты классификации / Виды деятельности	Количество сеансов тестовой выборки	Поиск информации на Web-серверах	Загрузка данных	Отправка электронной корреспонденции	Получение электронной корреспонденции	Отсутствует достоверный результат классификации
Поиск информации на Web-серверах	84	63	0	3	2	16
Загрузка данных	42	0	41	0	0	1
Отправка электронной корреспонденции	36	0	0	30	0	6
Получение электронной корреспонденции	44	0	0	0	40	4

Предложенный метод позволяет ввести функцию контроля за сетевой деятельностью пользователей, повышая возможность обеспечения безопасности вычислительных сетей. Простая реализация метода позволяет произвести легкое внедрение разработок в инструментарий специалиста безопасности при анализе и классификации деятельности пользователей.

Литература

1. Козьмовский Д.В., Куватов В.И., Пантиховский О.В. К вопросу о классификации деятельности пользователей в распределенных сетях: тр. XII С.-Петерб. междунар. конф. «Региональная информатика (РИ-2010)». СПб.: СПОИСУ, 2011. С. 157–160.

2. Козьмовский Д.В., Пантиховский О.В., Смирнов А.С. Проблемы безопасности информации распределенных информационных систем объектов управления // Проблемы управления рисками в техносфере. 2011. № 2 (18). С. 88–92.

3. Манита А.Д. Теория вероятностей и математическая статистика: учеб. пособие. М.: Изд. отдел УНЦ ДО, 2001. 256 с.

4. Козьмовский Д.В., Куватов В.И., Примакин А.И. Методы анализа трафика и определения сетевой деятельности в вычислительных сетях в интересах контроля пользователей // Вестник Санкт-Петербургского университета МВД России. 2014. № 1 (61). С. 112–115.

5. Козьмовский Д.В., Малыгин И.Г. Методы обеспечения безопасности распределенных информационных систем МЧС России, основанных на анализе трафика и контроле сетевой деятельности пользователей // Проблемы управления рисками в техносфере. 2013. № 2 (26). С. 78–82.