БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Insertion, Evasion, and Denial of Service: Eluding Network Intrusion Detection. Thomas H. Ptacek, Timothy N. Newsham [Электронный ресурс] / Режим доступа: http://insecure.org/stf/secnet_ids/secnet_ids.html, свободный. – Загл. с экрана.
2. Network Based Intrusion Detection. A review of technologies. [Электронный ресурс] / Режим доступа: http://linkinghub.elsevier.com/retrieve/pii/S01674048998 0131X, свободный. – Загл. с экрана.
3. Benchmarking network IDS. [Электронный ресурс] / Режим доступа: http://archives.neohapsis.com/ archives/sf/ids/2000-q4/0244.html, свободный. – Загл. с экрана.
4. Common Vulnerabilities and Exposures [Электронный ресурс] / Режим доступа: http://Cve.mitre.org, свободный. – Загл. с экрана.

Статью рекомендовал к опубликованию к.т.н., доцент О.Б. Спиридонов.

Абрамов Евгений Сергеевич
Технологический институт федерального государственного автономного образовательного учреждения высшего профессионального образования «Южный федеральный университет» в г. Таганроге.
E-mail: abramoves@gmail.com.
347928, г. Таганрог, пер. Некрасовский, 44.
Тел.: 88634371905.
Кафедра безопасности информационных технологий; к.т.н.; доцент.

Половко Иван Юрьевич
E-mail: ivan.polovko@mail.ru.
Кафедра безопасности информационных технологий; аспирант.

Abramov Evgeny Sergeevich
Taganrog Institute of Technology – Federal State-Owned Autonomy Educational Establishment of Higher Vocational Education “Southern Federal University”.
E-mail: abramoves@gmail.com.
44, Nekrasovskiy, Taganrog, 347928, Russia.
Phone: +78634371905.
The Department of Security in Data Processing Technologies; Cand. of Eng. Sc.; Associate Professor.

Polovko Ivan Yur’evich
E-mail: ivan.polovko@mail.ru.
The Department of Security in Data Processing Technologies; Postgraduate Student.

УДК 519.254, 004.056

В.А. Нестеренко, А.А. Таран

РЕДУКЦИЯ РАЗМЕРНОСТИ ПРОСТРАНСТВА СОСТОЯНИЙ В ЗАДАЧАХ АНАЛИЗА СЕТЕВОГО ТРАФИКА

Статья посвящена рассмотрению возможности уменьшения числа характеристик используемых при анализе состояния системы. Задача снижения числа характеристик очень важна при разработке и создании систем обнаружения вторжений: С увеличением числа характеристик улучшается качество систем обнаружения вторжений, с одной стороны, и уменьшается призводительность и быстродействие, с другой стороны. Рассмотрены два метода: метод главных компонент (principal component analysis – PCA) и линейный дискриминантный анализ Фишера (Fisher's linear discriminant analysis – LDA). Проводится оценка эффективности этих методов и примеры их практического использования при анализе сетевого трафика.

Метод главных компонент; линейный дискриминантный анализ; алгоритм Фишера; снижение размерности данных; обнаружение вторжений; анализ сетевого трафика.

V.A. Nesterenko, A.A. Taran

REDUCTION OF DIMENSION STATE-SPACE PROBLEM OF ANALYSIS OF NETWORK TRAFFIC

The article is devoted consideration of possibility of reduction of number of characteristics used at the analysis of a system. The problem of decrease in number of characteristics is very important by working out and creation of systems of intrusions detection: With increase in number of characteristics quality of systems of intrusions detection improves on the one hand and speed decreases on the other hand. Two methods are considered: Method of principal component analysis (PCA) and Fisher's linear discriminant analysis (LDA). The estimation of efficiency of these methods and examples of their practical use is spent at the analysis of the network traffic.

Method of principal component analysis – PCA; Fisher's linear discriminant analysis – LDA; decrease in dimension of the data; detection of intrusions; the analysis of the network traffic.

Введение. Обеспечение информационной безопасности в условиях высокой интенсивности передачи данных связано с проблемами обработки больших объемов информации. Так, в исследуемой сети или на локальном компьютере постоянно происходит множество событий (системные вызовы, открытие, копирование или удаление файлов, отправка или получение пакетов через сеть, нажатие определенных клавиш и т. п.). Каждое из них описывается с помощью нескольких десятков числовых и качественных характеристик (тип протокола, количество передаваемых байт, адрес источника). При этом количество характеристик должно быть достаточным, чтобы по ним можно было однозначно определить, является ли рассматриваемое событие нормальным для данной системы или это аномалия, возможно, вывзванная вредоносной активностью. Согласно стандарту RFC-1213 [1] для организации управления сетью используется 12 групп объектов, представители каждой из которых характеризуются примерно 20 параметрами. А организаторы соревнований KDDCup99 [2] использовали 41 параметр для описания сетевых соединений при поиске атак.

С другой стороны, при необоснованном росте размерности данных резко повышается ресурсоемкость, снижается быстродействие, а иногда и качество работы анализаторов. Это недопустимо для программного обеспечения, которое должно обрабатывать данные на лету, в режиме реального времени (системы обнаружения вторжений, брандмауэры, сетевые фильтры и т. п.). Поэтому часто системы обеспечения информационной безопасности в качестве первого шага своей работы применяют алгоритмы снижения размерности входных данных.

В предлагаемой статье рассматривается применение двух линейных алгоритмов снижения размерности в задачах классификации данных, полученных при анализе сетевого трафика. А именно, метод главных компонент (principal component analysis – PCA) и метод линейного дискриминантного анализа Фишера (Fisher's linear discriminant analysis – LDA). Далее приводится краткое описание обоих методов, результаты экспериментов по их приложению к задачам обнаружения сетевых вторжений и туннелирования, а также список достоинств и недостатков каждого алгоритма.

Редукция размерности. Редукция размерности с точки зрения интеллектуального анализа данных заключается в поиске оптимального представления данных в пространстве меньшей размерности с максимальным сохранением структуры и свойств исходного набора данных. При этом могут ставится несколько целей. Самая очевидная – устранение избыточности. Другая, более сложная – поиск представления, позволяющего максимально быстро определить основные свойства и структуру исходных данных.

Пусть $X \subset R^d$ – множество точек в пространстве характеристик, соответствующее множеству событий, возникающих в сети, каждое событие описывается

d характеритиками. Будем считать, что набор характеристик подобран так, что исходное множество однозначно разделяется на 2 различных множества Y и Z, соответствующих нормальным и аномальным событиям соответственно. Пусть при этом множество Y содержит N_Y элементов, а мощность множества Z – N_Z. При создании систем обнаружения нарушений необходимо решить задачу разбиения множества X на нормальные Y и аномальные Z события соответственно. Поскольку работать с большим количеством характеристик непосредственно неудобно (снижается скорость обработки данных, усложняются алгоритмы выявления структуры и т.д.), то для решения указаной задачи в первую очередь необходимо уменьшить размерность пространства характеристик без существенной потери информации о событиях в системе. Мы рассмотрим два линейных алгоритма снижения размерности: метод главных компонент и линейный дискриминантный алгоритм Фишера. В обеих случаях размерность входных данных снижается при помощи проектирования на выбранное линейное подпространство меньшей размерности. Пусть $k < d$ – размерность пространства проекций, тогда описаная операция может быть представлена как действие линейного отображения $\varphi : X \to R^k$ по правилу

$$\varphi : x \mapsto Cx = \begin{bmatrix} c_{11} & \dots & c_{d1} \\ & \dots & \\ c_{k1} & \dots & c_{kd} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \dots \\ x_d \end{bmatrix}.$$

В случае, когда необходимо разделить лишь на два множества (нормальные и аномальные события, например), наиболее удобным является проектирование в одномерное пространство. При удачном выборе проектируещей прямой классификация может быть осуществлена с использованием обычных операций сравнения (<,>,=) при подсчете разности значения проекции текущей точки и, например, проекций математических ожиданий исходных классов. Тогда действие оператора φ можно записать так

$$\varphi : x \mapsto c^\tau x = \begin{bmatrix} c_1 & \dots & c_d \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \dots \\ x_d \end{bmatrix}. \quad (1)$$

Способ выбора вектора *c* определяется используемым алгоритмом снижения размерности.

Метод главных компонент. Идея метода главных компонент состоит в выборе в исходном d-мерном пространстве характеристик ортонормированного базиса так, чтобы его первая компонента a_1 соответствовала направлению максимального разброса в пространстве событий, следующая a_2 определяла направление максимального рассеяния при вычитании из исходных данных проекций на первый базисный вектор и т.д. В конце концов будет получен набор векторов a_1, a_2,..., a_n. Из него выбираются первые k векторов. Они являются базисом линейного пространства, на которое будет осуществляться проектирование. Установленно [3], что заданный таким образом набор векторов представляет из себя множество собственных векторов, соответствующих собственным значениям ковариационной матрицы σ^2 множества событий X. В рассматриваемом случае требуется найти только вектор a_1, который соответствует максимально-

му собственному значению ковариационной матрицы σ^2 и направлению максимального разброса характеристик. И хотя метод главных компонент изначально не был предназначен для классификации данных, однако в случае, когда рассеяние элементов играет важную роль в разделении данных, может быть успешно применен для этой цели.

Метод главных компонент требует, чтобы исходный набор данных был центрированным набором данных. Поэтому первым шагом является поиск математического ожидания m множества X и покомпонентное вычитание его из всех элементов исходного множества. Следующий этап – рассчет ковариационной матрицы σ^2. Поскольку по свойствам ковариационной матрицы она симметрична и положительно определена, то для поиска a_1 может быть применен алгоритм прямых итераций [4]. Найденный собственный вектор a_1 будем использовать в качестве проекционного вектора c в формуле (1).

Из описания алгоритма ясно, что скорость его работы зависит от размерности исследуемых данных d, а весь алгоритм может быть разбит на две последовательные части: первая – этап обучения – включает в себя расчет математических ожиданий классов X, Y и их ковариационных матриц. Пусть для обучения системы используется n векторов, тогда для подсчета математического ожиданий требуется $d \cdot n$ сложений, а центрирование элементов требует столько же вычитаний. Для формирования ковариационной матрицы понадобится $n \cdot d^2$ сложений и $n \cdot d^2$ умножений. Количество операций, требуемых для поиска проектирующего вектора c, не может быть указано точно, поскольку для этого используется итерационный алгоритм, зависящий от заданной точности ε и выбора начального приближения x_0. Однако оно может быть ограничено заданным пользователем максимальным числом итераций I. Метод прямых итераций состоит в многократном умножении вектора x_0 на ковариационную матрицу σ^2. Поэтому для каждой итерации требуется d^2 умножений и d^2 сложений. Таким образом, общее количество операций: $2(n \cdot d + n \cdot d^2 + I \cdot d^2)$, а трудоёмкость работы алгоритма на обучающем этапе зависит от размерности данных и количества элементов обучающей выборки как $O(n \cdot d^2)$.

На втором этапе, этапе анализа данных о состоянии системы, производится центрирование и проектирование возникающих векторов на расчетную прямую. Это требует по d сложений, вычитаний и умножений на каждый входной вектор. То есть скорость работы алгоритма здесь зависит от числа использованных при сборе даных характеристик как $O(d)$.

Линейный дискриминантный анализ Фишера. Линейный дискриминантный анализ Фишера (LDA) был непосредственно разработан для решения задачи разбиения исходного набора данных на заданные множества наилучшим образом. Пусть m_Y и m_Z – векторы, соответствующие математическим ожиданиям исходных классов Y и Z соответственно, а внутриклассовые ковариационные матрицы для этих множеств – σ_Y^2 и σ_Z^2. Обозначим математические ожидания классов после проекций – m'_Y и m'_Z, а соответствующие дисперсии – σ'^2_Y и σ'^2_Z.

Фишер предложил [5], что лучшим будет такое направление проекций, при котором расстояние между центрами разделяемых классов окажется максимальным и в то же время разброс элементов в полученных проекциях множеств окажется минимальным. В связи с этим используется следующий критерий для поиска проектирующего вектора c (1):

$$J(c) = \frac{(m'_Y - m'_Z)^2}{\sigma'^2_1 + \sigma'^2_2} \to \max. \quad (2)$$

В этом случае для искомого вектора c можно получить следующее выражение:

$$c = \Omega_W^{-1}(m_Y - m_Z), \quad (3)$$

где $\Omega_W = \sigma_Y^2 + \sigma_Z^2$.

Как и в случае метода главных компонент, будем рассматривать количество выполняемых операций в зависимости от двух параметров: d – количество использованных характеристик для описания одного события во входном множестве X и n – объем обучающей выборки.

В процессе обучения подсчет математических ожиданий и ковариационнных матриц требует выполнения $2 \cdot d \cdot n$ сложений и $d \cdot n$ умножений. Подсчет матриц $m_Y - m_Z$ и Ω_W использует в сумме $d + d^2$ сложений. И наконец, расчет проектирующего вектора (3) равносилен решению следующей системы линейных уравнений:

$$\Omega_W c = (m_Y - m_Z).$$

Решение этой задачи, например методом Гаусса, зависит от размерности пространства характеристик как $O(d^3)$ [5]. Таким образом, ассимптотическая оценка трудоёмкости алгоритма на обучающем этапе составляет $O(n \cdot d + d^3)$.

Непосредственно проектирование требует d сложений и столько же умножений на каждый входной вектор. И так же как в методе главных компонент линейно ($O(d)$) зависит от размерности входного вектора.

Результаты практического применения к анализу трафика. При решении проблем информационной безопасности процесс анализа и классификации сетевого трафика разбивается на два класса задач: первый – поиск в наборе данных элементов, соответствующих заданным событиям, с помощью некоторого шаблона (метод сигнатур), второй – обнаружениие отклонений от заданной модели поведения (метод поиска аномалий). Первый тип позволяет быстро обнаруживать уже известные виды атак, другой – выявлять новые.

Поскольку метод главных компонент учитывает только общие свойства пространства характеристик и никак не использует информацию о множествах, по которым будет производиться классификация после сокращения размерности, то он не может быть использован в качестве элемента некоторого сигнатурного метода, однако может оказаться полезным при поиске аномалий.

С другой стороны, поиск проектирующего вектора алгоритмом Фишера жестко связан с математическими ожиданиями и ковариационными матрицами двух разделяемых классов, а значит, для корректной реализации этого метода требуется априорное знание структуры исследуемого пространства характеристик, что более соответствует определению сигнатурных методов анализа данных.

Эффективность применения этих алгоритмов для анализа реальных наборов данных можно увидеть на следующем примере. Имеется сборник icmp-пакетов, взятый из коллекции open-packets.org [6]. Часть из этих пакетов относится к атаке icmp-tunneling против машины с операционной системой Windows Server 2000, другая содержит набор пакетов echo-reply и echo-request, собранных на машинах с операционными системами Windows 7 и Linux Ubuntu 10.0. Для описания этих данных было использовано 20 характеристик, представляющих собой приведенные к десятичному виду пары байт поля «данные» этих пакетов.

Составив обучающие множества из 400 пакетов, относящихся к атаке, и 400 пакетов другого вида и применив к остальным данным метод Фишера на основе этих выборок, получаем безошибочное отделение атаки от нормального трафика. Этот результат иллюстрирует гистограмма (рис. 1).

При использовании той же обучающей выборки для метода главных компонент не удается отделить пакеты, отвечающие туннелированию, от соответствующих работе утилиты пинг (рис. 2): проекции обоих классов оказались рядом. Однако метод главных компонент позволяет увидеть другую структуру этого множества: в левой части гистограммы находятся проекции, отвечающие характеристикам icmp-пакетов, собранных на операционной системе Windows, в правой – Linux.

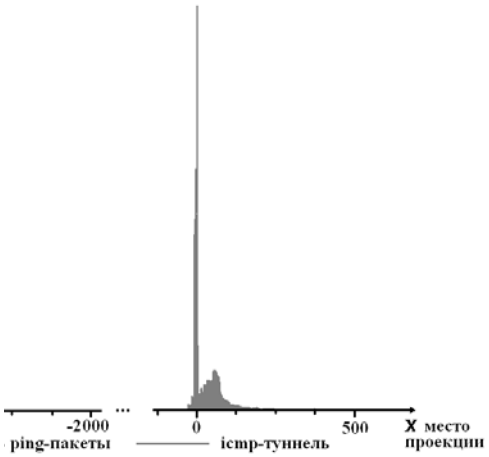

Рис. 1. Гистограмма для задачи обнаружения icmp-туннеля методом Фишера

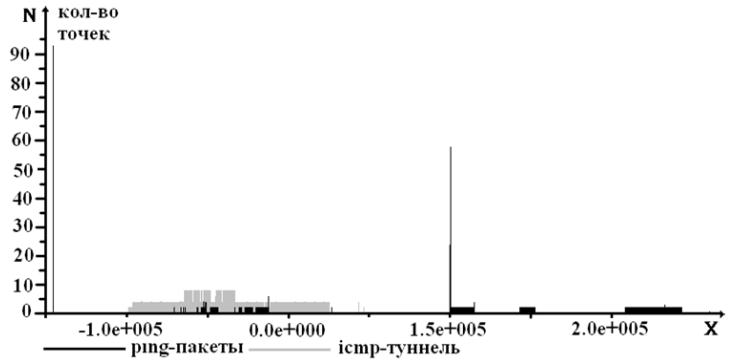

Рис. 2. Гистограмма проекций для задачи обнаружения icmp-туннеля методом главных компонент

Другой показательный пример – задача о разделении TCP-соединений, приходящихся на 80 (http) и 22 (telnet) порты. Для анализа были использованы пакеты, собранные в течение суток работы учебного класса мехмата ЮФУ. Каждому соединению были поставлены в соответствие 5 характеристик, а именно: продолжительность соединения, количество переданных пакетов от источника к приемнику и обратно, а также объем данных, переданных в обе стороны.

Поскольку каждое из анализируемых множеств в таком пространстве характеристик имеет плотную и однородную структуру и в то же время математические ожидания классов расположены на дастаточном расстоянии друг от друга, то применение обоих алгоритмов дает схожие результаты. При линейном дискриминантном анализе (рис. 3) из 4900 соединений лишь 7 были классифицированы не верно. В случае метода главных компонент количество ошибок несколько больше (100 элементов), однако множества отделимы друг от друга даже визуально (рис. 4).

На последнем примере можно увидеть еще одно интересное свойство проекционного вектора c. Рассмотрим полученный вектор. Его компоненты – коэффициенты в линейной комбинации характеристик. Наиболее значимыми для разделения событий на классы будут характеристики, соответствующие наибольшим по модулю компонентам проекционного вектора.

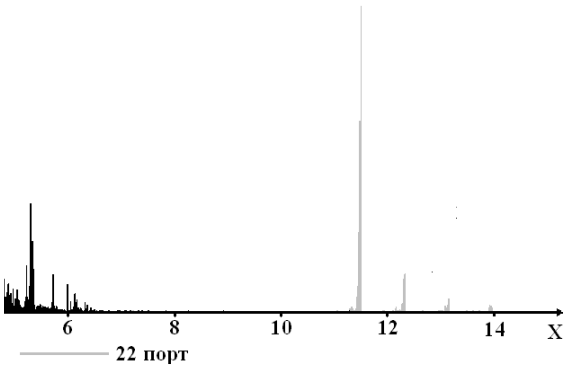

Рис. 3. Гистограмма проекции для задачи разделения соединений, проходящих через 80 и 22 порты методом Фишера

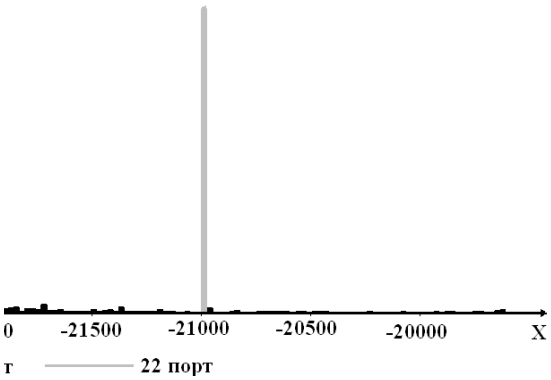

Рис. 4. Гистограмма проекции для задачи разделения соединений, проходящих через 80 и 22 порты методом главных компонент

Так, например, вектор проектирования, полученный при классификации соединений методом алгоритма Фишера, относящихся к 80 и 22 порту, имеет следующий вид:

$$c^{\tau} = (-0.040045, 0.826241, 0.424410, -0.000785, -001266).$$

Видно, что максимальные значения соответствуют второй и третьей характеристикам. В данном случае – это количество пакетов, переданных от исследуемого порта и к нему. Хотя этих двух параметров и не достаточно для окончательной классификации, однако они играют важную роль в процессе разделения двух классов событий. Такие результаты позволяют быстро выявлять наиболее значимые свойства множеств при исследовании данных. Это может оказаться полезным, например, при изучениии вновь обнаруженных (0-day) атак.

Заключение. Линейный дискриминантный анализ Фишера и метод главных компонент как представители класса алгоритмов снижения размерности являются отличными инструментами для аналитиков сетевого трафика. Они не только позволяют отбросить лишнюю информацию, взглянуть на входные данные под другим углом и, возможно, лучше понять их структуру, но и могут стать хорошим дополнением к системам обнаружения вторжений или брандмауэрам. Так, алгоритм Фишера позволяет реализовать сигнатурный подход. Причем это проявляется и в части упрощения поиска соответствий с шаблонами, и при выборе наиболее значимых компонент для построения самих сигнатур. Метод главных компонент, с другой стороны, успешно применяется к задаче поиска аномалий.

Оба алгоритма просты в реализации, а скорость их работы линейно зависит от объема входных данных, что позволяет использовать их в системах реального времени. Однако стоит помнить, что не все данные имеют линейную структуру, а, значит, при применении к ним линейных методов редукции размерности могут оказаться утерянными важные свойства рассматриваемого трафика. Поэтому наилучшим решением в защите информации всегда является комбинация нескольких различных по сути алгоритмов.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. RFC-1213. Management Information Base for Network Management of TCP/IP-based internets: MIB-II. Network working group, http://www.ietf.org/rfc/rfc1213.txt.
2. Данные, использованные на соревнованиях KDD CUP 99. http://sigkdd.org/kddcup/index.php?section=1999&method=info.
3. *Veksler O.* Лекции по курсу распознавание образов. Университет Western Ontario, 2004. http://www.csd.uwo.ca/~olga/courses/CS434_541a/Lectures.pdf.
4. *Богачев К.Ю.* Практикум на ЭВМ. Методы решения линейных систем и нахождения собственных значений. – М.: МГУ им. Ломоносова, 1998.
5. *Fisher R.A.* The use of Multiple Mesurements in Taxonomic Problems. Annals of Eugenics. – 1936. – Vol. 7, part II.
6. Коллекция вредоносного трафика в формате pcap. https://www.openpacket.org/capture/by_category?category=Malicious.

Статью рекомендовал к опубликованию д.ф.-м.н.; профессор В.С. Малышевский.

Нестеренко Виктор Александрович
Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Южный федеральный университет».
E-mail: neva@sfedu.ru.
344082, г. Ростов-на-Дону, ул. Тургеневская, 27/30, кв. 32.
Тел.: 88632625798.
Доцент кафедры информатики и вычислительного эксперимента.

Таран Анна Александровна
E-mail: Annie4ka@yandex.ru.
г. Ростов-на-Дону, ул. Добровольского, 36/2, кв. 115.
Тел.: +7515034220; 88632749704.
Студент.

Nesterenko Victor Aleksandrovich
Federal State-Owned Autonomy Educational Establishment of Higher Vocational Education "Southern Federal University".
E-mail: neva@sfedu.ru.
27/30, Tourgenevsky Street, Fl. 32, Rostov-on-Don, 344082, Russia.
Phone: +78632625798.
Senior Lecturer of Chair of Computer Science and Computing Experiment.

Taran Anna Alexandrovna
E-mail: Annie4ka@yandex.ru.
36/2, Dobrovolskogo Street, Russia, Fl. 115, Rostov-on-Don, Russia.
Phone: +79515034220; +78632749704.
Student.

УДК 004.056.5, 004.89

А.В. Никишова

АРХИТЕКТУРА ТИПОВОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ ЗАДАЧИ ОБНАРУЖЕНИЯ АТАК

Рассмотрены основные тенденции развития атак. Предложена модель системы обнаружения атак, учитывающая их. Данная система обнаружения атак реализует сбор информации на нескольких уровнях информационной системы и использует для анализа системы искусственного интеллекта (нейронные сети). По результатам анализа ряда информационных систем организаций Волгограда была предложена архитектура типовой информационной системы. На ее основе была сформирован состав многоагентной системы обнаружения атак и деление ее агентов на миры.

Атака; система обнаружения атак; нейронная сеть; интеллектуальный агент; многоагентная система; миры; принятие совместного решения.

A.V. Nikishova

TYPICAL INFORMATION SYSTEM ARCHITECTURE FOR INTRUSION DETECTION PROBLEM

Major trends of attack's development have been considered. Intrusion detection system's model that takes them into consideration has been suggested. This intrusion detection system gathers information in several levels of information system and use artificial intelligence system (neural network) for analysis. According to the analysis of several information systems of Volgograd typical information system architecture was suggested. On its basis multi-agent intrusion detection system's structure and partition its agents into worlds.

Attack; intrusion detection system; neural network; intelligent agent; multi-agent system; worlds; make a joint decision.

В связи с широким распространением сетей общего пользования все большее число компьютеров подвергается атакам. Согласно статистике «Лаборатории Касперского» за 2010 г., количество новых атакующих воздействий держится на уровне 2009 г. и остается высоким (рис. 1), а общее количество инцидентов продолжает увеличиваться. В 2010 г. общее число зафиксированных инцидентов типа атаки через Интернет и локальные инциденты превысило 1,9 млрд.