- подверженных атаке или уязвимости операционных системах и приложениях;
- примеры ложных срабатываний;
- дополнительные ссылки на патчи и обновления, устраняющие данные проблемы и т. д.

Література: ***1.*** *Голяка А. Таксономия, тенденции развития систем обнаружения вторжений. «Правове, нормативне та метрологічне забезпечення системи захисту інформації в Україні»* ***2.*** *http://www.iss.net/.* ***3.*** *Петренко С. А., Петренко А. А. Аудит безопасности Intranet. ДМК Пресс, 2002.* ***4.*** *Сопроводительная документация к программному продукту Site Protector.*

УДК 004.4

ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В СИСТЕМАХ ОБНАРУЖЕНИЯ ВТОРЖЕНИЙ

Александр Голяка
НБУ

Аннотация: **Представлена возможность применения методов интеллектуального анализа данных в системах обнаружения атак.**
Summary: **Possibility application of Data Mining methods in the intrusion detection systems is presented.**
Ключевые слова: **Системы обнаружения вторжений, Data Mining, метод опорных векторов.**

Несмотря на то, что симбиоз RBID и SBID систем [1] повышает возможности обнаружения атак, однако, описанных в статье [1, 2] проблем SBID систем это не решает. Главной и принципиальной проблемой SBID систем являются либо ошибки первого, либо второго рода в терминах математической статистики, в зависимости от выставленного порогового значения для сигнализации атаки. В связи с этим, очевидна недостаточность хорошо изученного аппарата статистических моделей при построении SBID систем. В настоящее время ищутся новые методы анализа данных в системах обнаружения вторжений (СОВ) и все больше внимания уделяется применению методов интеллектуального анализа данных (ИАД, Data Mining). ИАД – это процесс выявления значимых корреляций, образцов и тенденций в больших объемах данных. Встречаются такие определения термина Data Mining:

- это процесс выделения из данных неявной и неструктурированной информации и представления ее в виде, пригодном для использования;

- это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных, интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

К методам и алгоритмам Data Mining относятся следующие: искусственные нейронные сети, деревья решений, символьные правила, методы ближайшего соседа и k-ближайшего соседа, метод опорных векторов, байесовские сети, линейная регрессия, корреляционно-регрессионный анализ; иерархические методы кластерного анализа, неиерархические методы кластерного анализа, в том числе алгоритмы k-средних и k-медианы, методы поиска ассоциативных правил, в том числе алгоритм Apriori, метод ограниченного перебора, эволюционное программирование и генетические алгоритмы, разнообразные методы визуализации данных и множество других методов.

Data Mining может состоять из трех стадий.

1. Выявление закономерностей (осуществляется исследование набора данных с целью поиска скрытых закономерностей; также должна осуществляться валидация закономерностей, т. е. проверка их достоверности на части данных, которые не принимали участие в формировании закономерностей);
2. Использование выявленных закономерностей для предсказания неизвестных значений (обнаруженные закономерности используются непосредственно для прогнозирования – решаются задачи классификации и прогнозирования);
3. Анализ исключений; стадия предназначена для выявления и объяснения аномалий, найденных в закономерностях (анализируются исключения или аномалии, выявленные в найденных закономерностях).

Методы Data Mining

Статистические методы Data Mining классифицированы на четыре группы:

1. дескриптивный анализ и описание исходных данных;
2. анализ связей (корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ);
3. многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.);
4. анализ временных рядов (динамические модели и прогнозирование).

Кибернетические методы Data Mining:

1. искусственные нейронные сети (распознавание, кластеризация, прогноз);
2. эволюционное программирование (в т. ч. алгоритмы метода группового учета аргументов);
3. генетические алгоритмы (оптимизация);
4. ассоциативная память (поиск аналогов, прототипов);
5. нечеткая логика;
6. деревья решений;
7. системы обработки экспертных знаний.

Основная идея этих методов применительно к СОВ основывается на предположении о том, что активность пользователей и программ в системе может быть отслежена и построена ее математическая модель. Для прикладного применения в СОВ методы ИАД можно рассмотреть с двух позиций: методы обнаружения нарушений (misuse detection), которые строят модель атаки, а в процессе обнаружения используют ИАД методы классификации; и методы обнаружения аномалий (anomaly detection), которые строят модель нормальной активности, а в процессе обнаружения используют ИАД методы поиска исключений.

Формально, будем использовать следующую модель задачи классификации.

Ω - множество анализируемых объектов (в терминах предмета распознавания образов – пространство образов).

$\omega : \overline{\omega} \in \Omega$ – объект классификации (распознавания) – образ.

$g(\omega) : \Omega \to \mathrm{M}$, M = {1,2,…,m} – индикаторная функция, разбивающая пространство образов Ω на m непересекающихся классов $\Omega^1, \Omega^2, ..., \Omega^m$. Индикаторная функция неизвестна наблюдателю.

X – пространство наблюдений, воспринимаемых наблюдателем – пространство признаков (признак – некоторое количественное измерение объекта).

$x(\omega) : \Omega \to X$ – функция, ставящая в соответствие каждому объекту ω точку $x(\omega)$ в пространстве признаков. Вектор $x(\omega)$ – это образ объекта, воспринимаемый наблюдателем.

В пространстве признаков определены непересекающиеся множества точек, соответствующих образам одного класса.

$\hat{g}(x) : X \to M$ – решающее правило – оценка для $g(\omega)$ на основании $x(\omega)$, т. е. $\hat{g}(x) = \hat{g}(x(\omega))$. (Решающим правилом называют правило отнесения образа к одному из классов на основании его вектора признаков).

Пусть $x_j = x(\omega_j)$, j = 1,2,…,N – доступная наблюдателю информация о функциях $g(\omega)$ и $x(\omega)$ (сами эти функции наблюдателю неизвестны). Тогда - (g_j, x_j), j = 1,2,…,N – есть множество прецедентов (образов, правильная классификация которых известна).

Задача заключается в построении такого решающего правила $\hat{g}(x)$, чтобы распознавание проводилось с минимальным числом ошибок.

Обычный случай – считать пространство признаков евклидовым, т. е. $X=R^1$. Качество решающего правила измеряют частотой появления правильных решений. Обычно его оценивают, наделяя множество объектов Ω некоторой вероятностной мерой. Тогда задача записывается в виде

$$\min P\{g(x(\omega)) \neq g(\omega)\} . \quad (1)$$

В настоящей работе предлагается использовать метод опорных векторов (SVM – Support Vector Machine) в системах обнаружения атак в качестве метода ИАД.

Метод опорных векторов

Рассмотрим задачу классификации для 2-х классов X_1 и X_2 (векторы пространства $\Re^n$). Т. е. на практике располагаем исходными данными, характеризующими нормальную активность пользователей, и некоторые примеры атак. Будем считать, что эти классы не пересекаются. Тогда существует единичный вектор φ и

число *c*, такие, что $(x_1,\varphi) > c$ при $x_1 \in X_1$ и $(x_2,\varphi) < c$ при $x_2 \in X_2$. В таком случае говорят, что X_1 и X_2 разделимы гиперплоскостью.

Обозначим $c_1(\varphi) = \min_{x_1 \in X_1}(x_1,\varphi)$ и $c_2(\varphi) = \max_{x_2 \in X_2}(x_2,\varphi)$. Тогда $(x_1,\varphi) > c_1(\varphi)$ при $x_1 \in X_1$, а $(x_2,\varphi) > c_2(\varphi)$ при $x_2 \in X_2$. Если $c_1(\varphi) \geq c_2(\varphi)$, то гиперплоскость

$$(x_1,\varphi) = \frac{c_1(\varphi) + c_2(\varphi)}{2} \qquad (2)$$

разделяет X_1 и X_2. Существует множество разделяющих гиперплоскостей в силу непрерывности $c_1(\varphi)$ и $c_2(\varphi)$. Задача состоит в нахождении оптимальной разделяющей гиперплоскости, формально, соответствующей вектору $\varphi_{опт}$, при котором достигается максимум $\Pi(\varphi)$ (логично, что разделяющая гиперплоскость должна быть расположена максимально далеко от ближайших к ней точек обоих классов). Доказана теорема, что если 2 множества X_1 и X_2 разделимы гиперплоскостью, то оптимальная разделяющая гиперплоскость существует и единственна.

Однако на практике выборка редко является линейно разделимой. Поэтому можно применять следующий удобный в СОВ подход – осуществить переход от исходного пространства признаковых описаний объектов X к новому пространству H с помощью некоторого преобразования $\psi : X \to H$. Если выборка в X не противоречива и H имеет достаточно высокую размерность, то всегда найдется пространство, в котором она разделима. Пространство H называют спрямляющим.

Функция $K : X \times X \to \Re$ называется ядром, если она представима в виде $K(x_1,x_2) = \langle \psi(x_1), \psi(x_2) \rangle_H$ при некотором отображении $\psi : X \to H$, где H – пространство со скалярным произведением (пространство евклидово, в общем случае гильбертово).

В качестве варианта метода решения задачи, необязательно заниматься подбором отображения ψ и строить H, а достаточно только подобрать ядро (т. н. kernel function).

Основная идея метода SVM проиллюстрирована на рис. 1.

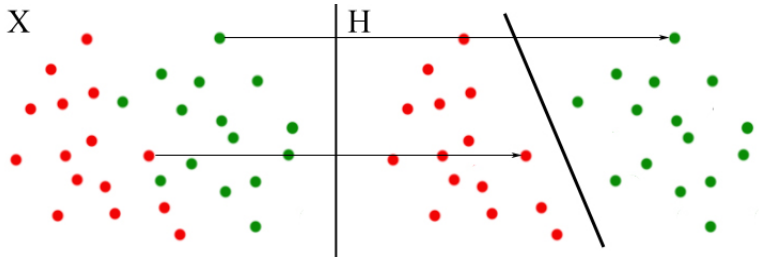

Рисунок 1 – Метод SVM

Исходные объекты (в левой части рисунка) преобразуются при помощи ядерных функций. После этого новый набор преобразованных объектов (в правой части рисунка) уже линейно разделим. Таким образом, вместо построения сложной кривой требуется лишь провести оптимальную прямую, которая отделит объекты типа GREEN от объектов типа RED (т.е. образы «нормальной» активности от атак).

В качестве ядерных функций предлагается использовать потенциальные функции (п. ф.). В этом случае, каждая точка (образ) образует в пространстве признаков *X* некоторое поле притяжения. Например, можно рассматривать каждую точку как точечный электрический заряд. Электрическое поле описывается потенциалом, создаваемым системой зарядов во всем пространстве. Изменение потенциала электрического поля по мере удаления от заряда обратно пропорционально квадрату расстояния. Потенциал, таким образом, может служить мерой удаления точки от заряда. Когда поле образовано несколькими зарядами, потенциал в каждой точке этого поля равен сумме потенциалов, создаваемых в этой точке каждым из зарядов. Если заряды, образующие поле, расположены компактной группой, потенциал поля будет иметь наибольшее значение внутри группы зарядов и убывать по мере удаления от нее. Тогда $K(x,y)$ – потенциальная функция, $x, y \in X$ такая, что:

$K(x,y) > 0$, при $x \neq y$,

$$K(x,y) = K(x, x+\mu(y-x)) = \tilde{K}(\mu), \tag{3}$$

где $\tilde{K}(\mu)$ - монотонно убывающая функция и $\tilde{K}(0)$ – ее максимальное значение.

При использовании указанных выше методов и в результате при построении оптимальной канонической гиперплоскости $\langle w, p \rangle$ в пространстве характеристик *H*, приходим к решению следующей оптимизационной задачи:

$$\min_{w\in H, \zeta_i \in \Re} \frac{1}{2}\| w \|^2 + \frac{1}{vl}\sum_i^l \zeta_i - 1, \tag{4}$$

при $(w \cdot \varphi(x_i) - p) \geq 1 - \zeta_i, \zeta_i \geq 0, \forall i \in [1, N]$.

После решения оптимизационной задачи, решающая функция для каждой точки x имеет вид:

$$f(x) = \operatorname{sgn}((w \cdot \varphi(x)) - 1) \tag{5}$$

Если использовать метод Лагранжа и ввести дополнительные переменные (множители Лагранжа) a_i, тогда можно представить оптимизационную задачу как

$$\min : \frac{1}{2}\sum_{i,j} a_i a_j K_\varphi(x_i, x_j), \text{ при } 0 \leq a_i \leq \frac{1}{vl}, \sum_i a_i = 1. \text{ А}$$

Решающая функция принимает вид:

$$f(x) = \operatorname{sgn}\left(\sum_i a_i K_\varphi(x_i, x) - 1 + p\right). \tag{6}$$

Параметр регуляризации *v* задает компромисс между точностью модели, определяемой величиной тренировочной ошибки $\sum \zeta_i$, и способностью модели к обобщению, определяемой величиной границы $\frac{1}{2}\| w \|^2$, $0 < v < 1$. Параметр устанавливается априори.

Достоинствами SVM с использованием п. ф. являются:

1. получение функции классификации с минимальным уровнем ошибки классификации;
2. возможность использования линейного классификатора для работы с нелинейно разделяемыми данными, сочетая простоту с эффективностью;
3. возможность работы с разнородными сложно структурированными данными за счет использования различных п. ф.;
4. в случае изменения структуры анализируемых данных, достаточно заменить только используемую п.ф., без замены самого алгоритма;
5. по сути в SVM решается главным образом задача квадратичного программирования, имеющая единственное решение, и для нее существует множество изученных эффективных методов оптимизации, что позволяет работать в режиме реального времени.

Однако SVM имеет и некоторые незначительные недостатки, а именно:

1. решающая функция *f*(*x*) зависит от параметра *v*, устанавливаемого априори;
2. SVM чувствителен к наличию «шума» в тренировочном наборе.

Для преодоления этих недостатков в качестве одного из вариантов предлагается использовать математический аппарат нечетких множеств.

Подход к формализации понятия нечеткого множества состоит в обобщении понятия принадлежности. В обычной теории множеств существует несколько способов задания множества. Одним из них является задание с помощью характеристической функции, определяемой следующим образом. Пусть U – так называемое универсальное множество, из элементов которого образованы все остальные множества, рассматриваемые в данном классе задач, например множество всех целых чисел, множество всех гладких функций и т. д. Характеристическая функция множества $A \subseteq U$ – это функция μ_A, значения которой указывают, является ли $x \in U$ элементом множества *A*:

$$\mu_A(x) = \begin{cases} 1, x \in A, \\ 0, x \notin A \end{cases} \tag{7}$$

Особенностью этой функции является бинарный характер ее значений.

С точки зрения характеристической функции нечеткие множества есть естественное обобщение обычных множеств, когда мы отказываемся от бинарного характера этой функции и предполагаем, что она может принимать любые значения на отрезке [0,1]. В теории нечетких множеств характеристическая функция называется функцией. принадлежности, а ее значение $\mu_A(x)$ – степенью принадлежности элемента *x* нечеткому множеству *A*.

Более строго, нечетким множеством *A* называется совокупность пар

$$A = \{\langle x, \mu_A(x)\rangle \mid x \in U\},$$

где μ_A – функция принадлежности, т. е. $\mu_A : U \to [0,1]$.

Итого, в нашу оптимизационную задачу необходимо включить нечеткую функцию принадлежности элементов тренировочного набора $\mu_A(x)$. В результате «шумы» будут иметь меньшую степень принадлежности, чем корректные значения, и суть задачи в необходимости построения гиперплоскости *H*, разделяющей два нечетких множества.

В работе рассматривались вопросы применения методов интеллектуального анализа данных (Data Mining) для одной из задач обеспечения компьютерной безопасности – задачи выявления вторжений в компьютерные системы. Поскольку традиционные сигнатурные методы не обеспечивают должного уровня защиты, использование Data Mining методов в СОВ является активно развивающимся направлением. Основное внимание в работе уделено методам анализа данных на основе потенциальных функций, которые перспективны для данного направления.

Информация, которая может быть основой построения описанных методов, может быть подана в разных формах, например, в виде трудно объяснимых проблем в компьютерных системах, диапазонов пороговых значений, параметров входного/выходного трафика, непредусмотренных адресов пакетов, атрибутов, временных параметров, запросов и т. д.

Однако следует помнить, что реализация описанной методики построения СОВ вряд ли заменит промышленную полнофункциональную систему, с многолетним опытом работы на рынке безопасности. В цикле статей [1, 2] и данной статье предлагается лишь дополнить функциональность системы SiteProtector дополнительным эвристическим модулем обнаружения атак на случай появления атаки, по каким либо причинам не внесенной в базу данных сигнатур атак.

Література: ***1.*** *Голяка А. Таксономия, тенденции развития систем обнаружения вторжений. «Правове, нормативне та метрологічне забезпечення системи захисту інформації в Україні»* ***2.*** *Голяка А. Архитектура, взаимодействие компонентов, преимущества siteprotector. «Правове, нормативне та метрологічне забезпечення системи захисту інформації в Україні»* ***3.*** *Eleazar Eskin Christina Leslie and William Stafford Noble. The spectrum kernel: A string kernel for SVM protein classification. In Proceedings of the Pacific Symposium on Biocomputing (PSB-2002), Kaua'I, Hawaii, 2002.* ***4.*** *N. Cristianini and J. Share-Taylor. An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK, 2000.* ***5.*** *D. E. Denning. An intrusion detection model. IEEE Transactions o Software Engineering, SE-13:222-232, 1987.* ***6.*** *W. Fan and S. Stolfo. Ensemble-based adaptive intrusion detection. In Proceeding of 2002 SIAM International Conference on Data Mining, Arlington, VA, 2002.* ***7.*** *K. Muller, S. Mika, G. Ratsch, K. Tsuda, B. Scholkopf, "An Introduction to Kernel-Based Learning Algorithms," IEEE Neural Networks, 12(2):181-201, May 2001.* ***8.*** *М. А. Айзерман, Э. М. Браверман, Л. И. Розоноэр. Метод потенциальных функций в теории обучения машин. Наука, Москва, 1970, 384 с.*

УДК 638.322

РЕАЛІЗАЦІЯ ПОВНИХ ПІДСТАНОВОК ЗА ДОПОМОГОЮ БАГАТОМОДУЛЬНОГО КАСКАДУ НАЙПРОСТІШИХ КОНСТРУКТИВНИХ МОДУЛІВ

Володимир Тарасенко, Олександр Тесленко, Олена Яновська
НТУУ «КПІ»

Анотація: **Досліджується можливість реалізації масових повних підстановок за допомогою найпростішого одновимірного каскаду конструктивних модулів.**