УДК 53.02+519.7+519.812.22

Е. Е. Витяев [1, 2], Б. Я. Ковалерчук [5], А. М. Федотов [1, 3], В. Б. Барахнин [1, 3], Д. С. Дурдин [1], С. Д. Белов [3], А. В. Демин [4]

[1] Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

[2] Институт математики им. С. Л.Соболева СО РАН
пр. Акад. Коптюга, 4, Новосибирск, 630090, Россия

[3] Институт вычислительных технологий СО РАН
ул. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

[4] Институт систем информатики СО РАН
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

[5] Central Washington University, Ellensburg, WA 98926-7520, USA
E-mail: fedotov@nsu.ru

ОБНАРУЖЕНИЕ ЗАКОНОМЕРНОСТЕЙ И РАСПОЗНАВАНИЕ АНОМАЛЬНЫХ СОБЫТИЙ В ПОТОКЕ ДАННЫХ СЕТЕВОГО ТРАФИКА *

В работе излагается новый подход к выявлению аномальных событий в потоке сетевого трафика на основе предложенной авторами гибридной корреляции событий (ГКС). Основная идея ГКС состоит в том, что бы аномальные события искать как нарушения нормального течения событий. Нормальное течение событий определяется совокупностью высоковероятных закономерностей, обнаруженных на данных нормального течения событий и определяющих в определённом смысле закон нормального течения событий. Для обнаружения закономерностей использовалась программная система Discovery, которая реализует разработанный авторами реляционный подход к интеллектуальному анализу данных. Система была применена к анализу данных сетевого трафика сервера системы передачи данных СО РАН. Было обнаружено значимое изменение в закономерностях нормального и анормального течения событий.

Ключевые слова: интеллектуальный анализ данных, анализ редких событий, обнаружение жульничества, Аномальные события.

В настоящее время с совершенствованием вычислительной техники и программного обеспечения одной из серьезнейших проблем становится компьютерная безопасность, слежение за состоянием системы, поиск сбоев и аномалий в работе систем. При этом с усложнением техники и программ, совершенствованием технологий записи и хранения данных объемы информации резко возросли.

Для анализа сбоев и аномалий в работе систем применяются алгоритмы Data Mining технологии *интеллектуального анализа данных* (ИАД), которая является мультидисциплинарной областью, возникшей на стыке методов машинного обучения (machine learning), искусственного интеллекта, прикладной статистики, распознавания образов, теории баз данных и др.

Однако современные схемы организации атак и других противозаконных действий характеризуются большой сложностью и запутанностью. Для распознавания этих схем необходимо обнаруживать сложные комбинации сетевых транзакций и сопутствующих им фактов, вследствие чего обучающее множество надо формировать не по одиночным транзакциям, а

* Работа выполнена при финансовой поддержке РФФИ (проект № 08-07-00272а), Совета по грантам Президента РФ (НШ-335.2008.1) в рамках государственного контракта № 02.514.11.4079, интеграционных проектов СО РАН № 1 и 115.

по комбинациям двух-трех и более транзакций. Это приводит к «взрыву» числа анализируемых цепочек транзакций.

Традиционные методы ИАД не приспособлены для решения таких задач. Методы ИАД требуют наличия обучающего материала в виде положительных и отрицательных примеров. Трудность состоит в том, что почти никогда нет обучающих данных по компьютерным атакам, аномалиям и противозаконным действиям, а если и есть, то число положительных случаев ничтожно и не позволяет сделать сколь-нибудь достоверных заключений.

В результате возникла проблема обнаружения редких или дисбалансных закономерностей (число обучающих случаев несопоставимо меньше общего числа случаев). Обнаружение таких закономерностей требует развития новых технологий ИАД [1–5].

Для решения этих задач возникла новая область интеллектуального анализа данных, называемая о*бнаружением взаимосвязей* (ОВ) (Link Discovery). Речь идет об обнаружении взаимосвязей между несопоставимыми, редкими и дисбалансными явлениями среди большого количества потоков и источников данных.

Применительно к сетевому трафику возникает задача реализации эффективных алгоритмов слежения за параметрами системы, обнаружения отклонений, сбоев и аномального поведения сети. При этом создаваемые алгоритмы должны обладать следующими возможностями:

1) проверка показаний характеристик сети на соответствие найденным закономерностям в ходе мониторинга сети;

2) анализ и прогнозирование возможного поведения сети при изменении ее характеристик;

3) обнаружение аномального поведения в сети.

В данной задаче для нас критичны следующие характеристики анализирующей системы:

а) высокая вероятность распознавания аномалий, наличие которых возможно определить по данным сети, собираемыми программами-коллекторами;

б) минимизация вероятности ложных срабатываний;

в) возможность алгоритмов детерминировать причину аномалии или помочь в этом пользователю системы;

г) гибкость в настройке – чувствительность, критерии поиска закономерностей;

д) производительность – малый расход системных ресурсов на проверку выборки на наличие в ней аномалий;

е) производительность – время обучения экземпляра анализатора.

Для решения этих задач нами предлагается использовать *реляционный* подход к *обнаружению знаний* (РОЗ) (Relational Data Mining [6–9]). В рамках данного подхода нами разработана технология автоматического обнаружения атак на компьютеры и сервера, названная *гибридной корреляцией событий* (ГКС) [10–11; 7; 8]. Эта технология состоит в автоматическом извлечении знаний из совокупности данных путем обнаружения высоковероятных закономерностей и последующим обнаружением нарушений этих закономерностей. Мы исходим из предположения, что высоковероятные закономерности определяют нормальное течение событий, а нарушение закономерностей – аномальное: атаку, жульническое или вредоносное течение событий. Важной особенностью этого подхода является возможность обнаружения новых, не встречавшихся ранее атак. Для обнаружения новых атак не требуется наличия обучающего материала, т. е. наличия уже идентифицированного набора атак определенного рода, как это требуется для существующих методов ИАД.

Формально это означает, что высоковероятная закономерность (гипотеза) и ее нарушение имеют вид:

Высоковероятная закономерность (гипотеза) $P \Rightarrow Q$ течения событий с вероятностью (> 0,9…) означает нормальное (штатное) течение событий. (1)

Нарушение закономерности означает выполнение $P \Rightarrow$ не(Q), что является нарушением нормального течения событий (атаку).

Новизна подхода и метода

Во всех методах ИАД, таких как нейронные сети, байесовские сети, решающие деревья, методы классификации, методы регрессионного анализа, распознавания, обнаружения правил и т. д., обнаружение зависимостей осуществляется в один шаг:

$$C = (A \,\&\, B \,\&\ldots\& \, G \Rightarrow Q).$$

Наш подход – двухступенчатый:

1) обнаружить высоковероятную закономерность $C = (A \,\&\, B \,\&\ldots\& \, G \Rightarrow Q)$;

2) взять ее отрицание not $C \Rightarrow S$, где not $C = (A \,\&\, B \,\&\ldots\& \, G \Rightarrow \text{not } (Q))$.

Здесь S – подозрительная ситуация. Обобщение, получаемое методами машинного обучения, требует некоторого набора аномальных ситуаций. Наш подход не требует наличия аномальных ситуаций для обучения и, следовательно, может обнаруживать новые и неизвестные заранее аномальные ситуации.

Для обнаружения закономерностей нами предлагается использовать РОЗ и ГКС, которые успешно применялись для решения задач финансового прогнозирования, медицины, биоинформатики и обнаружения жульничества [6–8; 10–13].

Технология ГКС

Технология ГКС предполагает реализацию следующего метода анализа событий.

1. Сформулировать в виде класса гипотез нормальное течение событий. Для этого надо сначала взять данные, описывающие эти события, и выделить словарь описания этих событий. С этой целью можно воспользоваться технологиями Netflow, позволяющими собирать статистику. В терминах этого словаря можно формулировать различные классы гипотез о нормальном течении событий:

а) задать максимально широкий класс гипотез вида (1);

б) сформулировать некоторый класс гипотез, характеризующих нормальное течение интересующего нас класса событий; не обязательно задавать все детали событий, достаточно задать только самые характерные свойства, остальные детали и уточнения будут обнаружены автоматически в процессе *семантического вероятностного вывода* (СВВ, см. определение ниже);

в) взять любое конкретное событие или шаблон и записать его в выделенных терминах и обобщить, заменив имена и константы переменными.

2. Обнаружить в классе проверяемых гипотез (1) те из них, которые являются высоковероятными закономерностями. Для обнаружения высоковероятных закономерностей используется семантический вероятностный вывод, который автоматически добавляет в гипотезу дополнительные признаки, обеспечивающие максимальное усиление закономерности и приближение ее вероятности к 1. В результате будут найдены все возможные уточнения гипотез, максимально точно описывающих штатную ситуацию. Доказано [9; 14–15], что СВВ обнаруживает все высоковероятные закономерности и тем самым решает задачу полного описания штатных ситуаций.

3. Получить множество всех *аномальных ситуаций* (АС), взяв отрицания обнаруженных высоковероятных закономерностей; использовать множество АС для обнаружения атак, жульничества или аномальных ситуаций.

4. Если в процессе применения множества АС обнаружена аномальная ситуация S, то о ней надо сигнализировать администратору сети. Администратор должен проверить действительно ли ситуация является аномальной. Если да, то принять соответствующие меры и затем определить, следствием чего она явилась, и включить дополнительные аномальные признаки ситуации в описание ситуации в АС. Если нет, то найти в ситуации S дополнительные параметры, которые определяют ее нормальность, и ввести отрицание этих параметров в описание ситуации S в АС. Это исключит ее применение в подобных ситуациях.

Ниже приведена схема алгоритма обнаружения анормальных событий.

Выделить словарь описания нормальных и аномальных событий и сгенерировать множество предикатов $\mathrm{P} = \{P_1, P_2, \ldots, P_m\}$ и предложений $A_1, A_2, \ldots, A_n$, формализующих эти описания в языке логики первого порядка.

Сформулировать класс гипотез $\{A_1 \& A_2 \& \ldots \& A_{n-1} \Rightarrow A_n\}$, описывающих нормальное течение событий.

Тестировать статистическую значимость – «существенность» каждого из условий $A_1, A_2, \ldots, A_{n-1}$ в гипотезе $(A_1 \& A_2 \& \ldots \& A_{n-1} \Rightarrow \mathrm{A_n})$ с помощью критерия Фишера [16]. Если ги-

потеза проходит критерий Фишера с определенным уровнем значимости для каждого условия $A_1, A_2, \ldots, A_{n-1}$ гипотезы, то эта гипотеза становится претендентом на высоковероятную закономерность.

Вычислить *условную частоту* (УЧ) гипотезы $(A_1\&A_2\&\ldots\&A_{n-1} \Rightarrow A_n)$ на данных.

$$\text{УЧ} = P(A_n / A_1\&A_2\&\ldots\&A_{n-1}) = N(A_n / A_1\&A_2\&\ldots\&A_{n-1}) / N(A_1\&A_2\&\ldots\&A_{n-1}\&A_n),$$

где $N(A_n / A_1\&A_2\&\ldots\&A_{n-1})$, $N(A_1\&A_2\&\ldots\&A_{n-1})$ – числа событий $A_1\&A_2\&\ldots\&A_{n-1}\&A_n$ и $A_1\&A_2\&\ldots\&A_{n-1}$.

Задать порог $T = 0{,}9$ для УЧ. Если $P(A_1\&A_2\&\ldots\&A_{n-1} \Rightarrow A_n) > T$, то гипотеза проходит еще один критерий и считается высоковероятной закономерностью, описывающей «нормальную» ситуацию.

Взять отрицание высоковероятной закономерности $A_1\&A_2\&\ldots\&A_{n-1} \Rightarrow \neg A_n$

Вычислить ее вероятность: $P(A_1\&A_2\&\ldots\&A_{n-1} \Rightarrow \neg A_n) = 1 - P(A_1\&A_2\&\ldots\&A_{n-1} \Rightarrow A_n)$.

Проанализировать случаи, которые описывают аномальную ситуацию

$$A_1\&A_2\&\ldots\&A_{n-1}\&\neg A_n.$$

Следовать рекомендациям предыдущего описания (п. 4) для уточнения ситуации и занесения ее во множество АС.

Описание алгоритма

1. *Форма представления закономерностей.* Предлагаемый нами метод оперирует закономерностями в виде *правил*:

$$A_1\&\ldots\&A_n \to B,$$

где $A_1, \ldots, A_n, B$ – предикаты, B называется *целевой* (THEN) частью, $A_1\&\ldots\&A_n$ называется *условной* (IF) частью.

Предполагается, что интервалы значений каждого признака определены заранее или получены некоторым алгоритмом кластеризации. Интервалы не должны пересекаться и должны покрывать все допустимые значения признака. Эти интервалы назовем *атомарными интервалами*, а соответствующий *предикат – атомарным*. Очевидно, что для представления множества всех логических формул, выражающих зависимости между значениями наблюдаемых признаков, нужно ввести логическую операцию отрицания, чтобы формулы логики первого порядка, не содержащие кванторов, могли быть представлены наборами правил. Мы предлагаем иное решение. Так как атомарные интервалы покрывают множество допустимых значений признака, то отрицание атомарного интервала эквивалентно принадлежности значения признака одному из оставшихся атомарных интервалов. Поэтому мы используем предикаты истинные тогда и только тогда, когда значение признака принадлежит некоторому набору атомарных интервалов. Эти наборы атомарных интервалов назовем *комбинированными интервалами* и соответствующие *предикаты – комбинированными*. Очевидно, комбинированные интервалы могут пересекаться или даже целиком покрываться другими комбинированными интервалами. Это, в частности, позволяет использовать интервалы различной ширины в целевой части правил, что очень полезно, когда нельзя сделать однозначный прогноз, в какой из атомарных интервалов попадает значение признака.

2. *Общая схема работы алгоритма поиска закономерностей.* Алгоритм поиска закономерностей основывается на методологии семантического вероятностного вывода [9; 14], позволяющего находить все максимально специфические и максимально вероятные закономерности в данных [9]. Определим на высказываниях языка первого порядка вероятность, как описано в [17]. Под семантическим вероятностным выводом понимается такая последовательность правил $C_1, C_2, \ldots, C_n$, что:

1) $C_i = (A_1^i\&\ldots\&A_{k_i}^i \Rightarrow G)$, $i = 1, \ldots, n$;

2) C_i – *подправило* правила C_{i+1}, т. е. $\{A_1^i, ..., A_{k_i}^i\} \subset \{A_1^{i+1}, ..., A_{k_{i+1}}^{i+1}\}$;

3) $\text{Prob}(C_i) < \text{Prob}(C_{i+1})$, $i = 1, 2, ..., n-1$, где $\text{Prob}(C_i) = \text{Prob}(G/A_1^i \& ... \& A_{k_i}^i)$ – *условная вероятность* (УВ) правила

$$\text{Prob}(G/A_1^i \& ... \& A_{k_i}^i) = \text{Prob}(G \& A_1^i \& ... \& A_{k_i}^i) / \text{Prob}(A_1^i \& ... \& A_{k_i}^i);$$

4) C_i – *вероятностные законы* (ВЗ), т. е. для любого подправила $C' = (A_1 \& ... \& A_j \Rightarrow G)$ правила C_i, $\{A_1, ..., A_j\} \subset \{A_1^i, ..., A_{k_i}^i\}$ выполнено неравенство $\text{Prob}(C') < \text{Prob}(C_i)$;

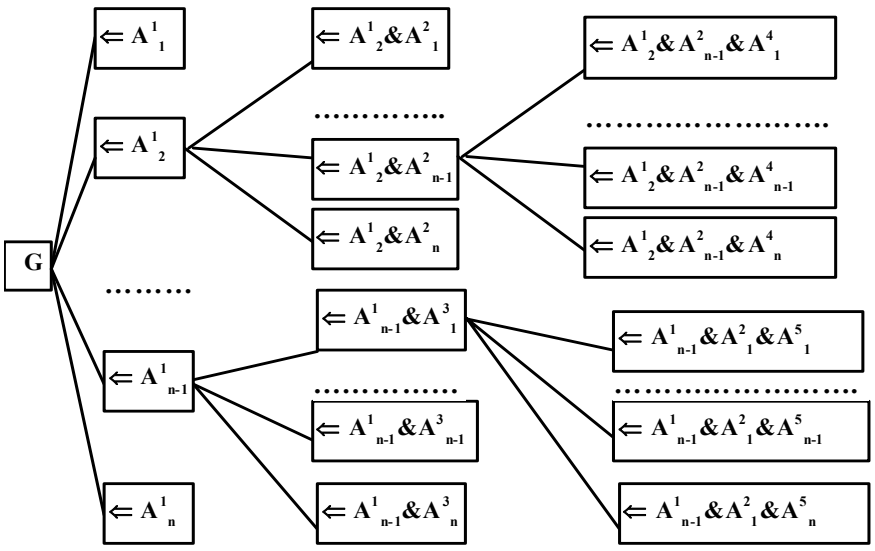
5) C_n – с*ильнейший вероятностный закон* (СВЗ), т. е. правило C_n не является подправилом никакого другого вероятностного закона.

Вероятностные неравенства в п. 3–4 проверяются на данных с использованием точного критерия независимости Фишера.

Применять методологию СВВ в чистом виде не представляется возможным в первую очередь из-за требований к производительности, а также из-за ограниченности выборки, на которой производится обучение. Следует применять следующие эвристики.

Первая эвристика состоит в том, что при построении СВВ алгоритм всегда добавляет к его условной части только один предикат. Как показывают некоторые предварительные оценки и эксперименты, крайне редка ситуация, когда добавление в условную часть правила сразу двух предикатов дает ВЗ, а добавление любого из этих двух признаков не дает ВЗ. Следовательно, данная эвристика почти не снижает количество и качество извлеченных из данных закономерностей при серьезном уменьшении пространства перебора.

Второй эвристикой является многоуровневая схема генерации правил. Схема состоит в следующем (дерево СВВ, включающее все СВВ, содержащие в заключении атом G):



Сначала находим все ВЗ с одним предикатом в условной части и заключением G, затем с двумя, тремя и т. д. При проверке, является ли правило ВЗ, проверяются только условные вероятности подправил предыдущего уровня дерева СВВ.

Другим преимуществом данной схемы является то, что она позволяет нам не хранить статистику для правил предыдущих уровней, что существенно сокращает затраты оперативной памяти на этапе генерации правил.

Как показало предварительное тестирование алгоритма, подсчет статистики для каждого правила является критичным для производительности программы. Поэтому для ускорения счета был проведен ряд алгоритмических и технических оптимизаций. Статистика для каждого предиката, условной и целевой частей правила представляется в виде массива чисел,

разрядность которых совпадает с разрядностью процессора. Операции над статистикой проводятся с помощью побитовых логических операций.

Пороговые величины: условная частота правила, уровень значимости критерия Фишера, максимальное число интервалов значений для признака и др., – могут задаваться пользователем перед началом обучения. Также учитываются атрибуты наблюдаемых признаков: Input, Predict и PredictOnly, которые указывают, в какой части правил (условной или целевой) находится предикат – только в условной, в обеих или только в целевой.

Результатом работы алгоритма являются:

- дерево СВВ для каждого целевого предиката;
- множество ВЗ и СВЗ этих деревьев;
- *максимально специфический закон* (МСЗ) для каждого целевого предиката, определяемый как СВЗ, обладающий наибольшей условной вероятностью среди других СВЗ дерева вывода этого предиката.

Множество всех МСЗ обладает таким важным свойством, как потенциальная непротиворечивость [9].

4. *Прогнозирование.* Для прогнозирования пользователь (эксперт, аналитик) определяет набор целевых признаков, для которых будет осуществляться предсказание, и набор признаков, по которым будет осуществляться предсказание.

Чтобы спрогнозировать наиболее вероятные значения целевых признаков, система должна применить найденные МСЗ к входным данным. Наиболее вероятное значение целевого признака будет определено как середина интервала, предсказываемого некоторым максимально специфическим правилом, для которого условная часть правила должна быть *выполнена на данных*, т. е. интервалы всех предикатов посылки правила должны накрывать значения соответствующих признаков в данных. В этом случае значение предсказываемого признака определяется однозначно.

В более сложном случае, когда максимально специфическое правило неприменимо к имеющимся данным, рассматриваются все ВЗ, выполнимые на данных. Из них выбираются все правила, имеющие в целевой части атомарные предикаты. Среди этих правил находится правило с максимальным значением УЧ.

В случае, когда однозначный выбор максимально вероятного правила сделать трудно, следует брать правила с комбинированным целевым предикатом, состоящим из 2-х подряд идущих атомов. Если и в этом случае нет однозначного максимально вероятного правила, то надо брать правила, состоящие из 3-х и т. д. подряд идущих интервалов. Нужно расширять интервал целевого предиката с целью увеличения УЧ предсказания. Для выбора максимально вероятного правила используется специальная оценочная функция, которая по набору лучших правил для каждого значения признака и статистики для этих правил определяет, однозначно ли можно выбрать доминирующее правило, по которому прогнозируется значение.

В случае, когда результатом прогноза должно быть одно значение, оно определяется как середина интервала, который предсказывается доминирующим правилом.

В случае же, когда результатом прогноза должна быть гистограмма, алгоритм выделят наиболее характерные интервалы значений по набору наиболее подходящих правил и возвращает статистическое распределение признака.

Результаты тестирования на данных сетевой статистики

В качестве анализируемых данных используются данные статистики, полученной с сервера, установленного в узле системы передачи данных СО РАН и предназначенного для раннего обнаружения вредоносных воздействий на сеть извне, проявлений аномального поведения компьютеров абонентов сети с целью обеспечения безопасности сети в целом. Для анализа использовалась статистика за «чистый» период, когда сеть функционировала нормально, и за аномальный период, когда в сети наблюдалась аномальная активность. Аномальный период характеризовался тем, что с определенного хоста посылалось большое количество ICMP-пакетов другим хостам сети СО РАН и во внешнюю сеть, что составляло около 80 % всего внешнего трафика сети СО РАН.

После предварительной обработки данные были собраны в одну таблицу, состоящую примерно из 750 000 записей. Каждая запись описывает событие пересылки данных между хостами по определенному протоколу за пятиминутный интервал времени и содержит 7 признаков. Около 450 000 записей в таблице соответствуют «чистому» периоду и 300 000 – аномальному периоду: всем пакетам, зарегистрированным за 5-минутные интервалы времени соответственно 18 февраля 2008 г. с 10.30 до 10.35 и 25 февраля 2008 г. с 10.30 до 10.35. Транзитный трафик не учитывался.

Анализируемые признаки:

Protocol ID – идентификатор протокола (номинальный тип данных);

Net ID Src, Net ID Dest – идентификаторы сетей, которым принадлежат IP-адреса отправителя и получателя соответственно (номинальный тип данных).

Адресное пространство IP-адресов разбито на следующие сети:

84.237.0.0/17 – часть адресного пространства сети СО РАН;
193.124.35.0/24 – часть адресного пространства сети СО РАН;
193.124.39.0/24 – часть адресного пространства сети СО РАН;
193.124.169.0/24 – часть адресного пространства сети СО РАН;
212.192.160.0/19 – часть адресного пространства сети СО РАН;
217.79.48.0/20 – часть адресного пространства сети СО РАН;
193.124.243.0/24 – часть адресного пространства сети СО РАН;
193.125.40.0/23 – часть адресного пространства сети СО РАН;
193.125.178.0/24 – часть адресного пространства сети СО РАН;
193.125.179.0/24 – часть адресного пространства сети СО РАН;
193.125.180.0/24 – часть адресного пространства сети СО РАН;
194.85.124.0/23 – часть адресного пространства сети СО РАН;
194.85.127.0/24 – часть адресного пространства сети СО РАН;
194.226.160.0/20 – часть адресного пространства сети СО РАН;
194.226.176.0/20 – часть адресного пространства сети СО РАН;
external – внешняя сеть;
multicast – широковещание.

Port Src, Port Dest – номера портов отправителя и получателя соответственно. Целочисленный, кластеризуемый тип данных. Если порт не определен для некоторого протокола, то его значение соответствует специальному кластеру со значением null.

Data Size – суммарный размер пакетов с уникальным набором пяти перечисленных выше признаков, зарегистрированных за 5-минутный интервал времени. Целочисленный, кластеризуемый тип данных.

Anomalous – признак, характеризующий период, в который была получена данная запись в таблице. Единица означает аномальный период (25 февраля), ноль означает нормальный период (18 февраля). Номинальный тип данных.

Для анализа был применен анализирующий модуль, созданный на основе системы Discovery, работающий под управлением Microsoft SQL Server 2005. Данный модуль разрабатывается для решения задач, связанных с извлечением знаний из различных источников данных, в том числе реляционных таблиц и многомерных моделей БД, и предоставления data mining функций другим приложениям.

В качестве целевого признака использовался признак «Anomalous». В качестве критериев статистической значимости, использовались точный критерий Фишера с пороговым значением 0,05 и критерий Юла с пороговым значением 0,1.

Были проведены следующие тесты

1. *Обнаружение вероятностных законов.* Ниже представлены СВЗ с условной вероятностью больше 0,7, обнаруженные системой Discovery (первый столбец – УЧ, второй – текстовое представление правила):

0,705	IF (28 байт ≤ Количество данных ≤ 198 байт) AND (Сеть получателя = 193.124.35.0/24) AND (4376 < Порт отправителя ≤ 41329) THEN (Аномальный период)
0,708	IF (Сеть отправителя = external network) AND (1 < Порт отправителя ≤ 65535) THEN («Чистый» период)
0,708	IF (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 65535) THEN («Чистый» период)
0,718	IF (Сеть отправителя = external network) AND (5820 < Порт получателя ≤ 42445) AND (Протокол = 17) THEN (Аномальный период)
0,729	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 55333) THEN («Чистый» период)
0,732	IF (96 байт < Количество данных ≤ 583 байт) AND (Сеть получателя = 84.237.0.0/17) AND (5820 < Порт получателя ≤ 42445) AND (4376 < Порт отправителя ≤ 41329) AND (Протокол = 17) THEN (Аномальный период)
0,774	IF (198 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 55333) AND (Протокол = 6) THEN («Чистый» период)
0,790	IF (28 байт ≤ Количество данных ≤ 1204 байт) AND (Сеть получателя = 84.237.0.0/17) AND (Протокол = 1) THEN (Аномальный период)
0,792	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть получателя = 194.226.160.0/20) THEN («Чистый» период)
0,794	IF (28 байт ≤ Количество данных ≤ 583 байт) AND (Сеть получателя = 84.237.0.0/17) AND (Протокол = 1) THEN (Аномальный период)
0,796	IF (198 байт < Количество данных ≤ 157746380 байт) AND (Сеть получателя = 194.226.160.0/20) AND (Протокол = 6) THEN («Чистый» период)
0,803	IF (583 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 55333) AND (Протокол = 6) THEN («Чистый» период)
0,807	IF (583 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт отправителя ≤ 65535) THEN («Чистый» период)
0,807	IF (583 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 65535) THEN («Чистый» период)
0,814	IF (583 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 55333) THEN («Чистый» период)
0,814	IF (28 байт ≤ Количество данных ≤ 198 байт) AND (Сеть получателя = 84.237.0.0/17) AND (Протокол = 1) THEN (Аномальный период)
0,826	IF (583 байт < Количество данных ≤ 157746380 байт) AND (Сеть получателя = 194.226.160.0/20) AND (Протокол = 6) THEN («Чистый» период)
0,826	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть получателя = 193.124.169.0/24) AND (1 < Порт получателя ≤ 65535) THEN («Чистый» период)
0,826	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть получателя = 193.124.169.0/24) AND (1 < Порт отправителя ≤ 65535) THEN («Чистый» период)
0,827	IF (583 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) THEN («Чистый» период)

0,827	IF (198 байт < Количество данных ≤ 157746380 байт) AND (Сеть получателя = 193.124.169.0/24) AND (1 < Порт отправителя ≤ 54928) THEN («Чистый» период)
0,828	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть получателя = 193.124.169.0/24) AND (1 < Порт отправителя ≤ 54928) THEN («Чистый» период)
0,840	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт отправителя ≤ 65535) THEN («Чистый» период)
0,840	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 65535) THEN («Чистый» период)
0,850	IF (96 байт < Количество данных ≤ 583 байт) AND (Сеть получателя = 84.237.0.0/17) AND (Протокол = 1) THEN (Аномальный период)
0,857	IF (28 байт ≤ Количество данных ≤ 583 байт) AND (Сеть отправителя = external network) AND (42445 < Порт получателя ≤ 65535) THEN (Аномальный период)
0,861	IF (583 байт < Количество данных ≤ 157746380 байт) AND (Сеть получателя = 193.124.169.0/24) AND (1 < Порт отправителя ≤ 54928) AND (Протокол = 6) THEN («Чистый» период)
0,863	IF (583 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) THEN («Чистый» период)
0,867	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт отправителя ≤ 54928) THEN («Чистый» период)
0,867	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт отправителя ≤ 65535) THEN («Чистый» период)
0,867	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 65535) THEN («Чистый» период)
0,868	IF (198 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (Протокол = 6) THEN («Чистый» период)
0,868	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 55333) THEN («Чистый» период)
0,870	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (Протокол = 6) THEN («Чистый» период)
0,870	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть получателя = 194.85.127.0/24) AND (Протокол = 6) THEN («Чистый» период)
0,870	IF (198 байт < Количество данных ≤ 157746380 байт) AND (Сеть получателя = 194.85.127.0/24) AND (Протокол = 6) THEN («Чистый» период)
0,870	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 65535) THEN («Чистый» период)
0,870	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт отправителя ≤ 65535) THEN («Чистый» период)
0,872	IF (583 байт < Количество данных ≤ 157746380 байт) AND (Сеть получателя = 194.85.127.0/24) AND (Протокол = 6) THEN («Чистый» период)
0,877	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 55333) AND (Протокол = 6) THEN («Чистый» период)

0,881	IF (198 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 55333) THEN («Чистый» период)
0,899	IF (96 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 55333) AND (Протокол = 6) THEN («Чистый» период)
0,903	IF (583 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт отправителя ≤ 54928) THEN («Чистый» период)
0,912	IF (583 байт < Количество данных ≤ 157746380 байт) AND (Сеть отправителя = external network) AND (1 < Порт получателя ≤ 55333) AND (Протокол = 6) THEN («Чистый» период)
0,933	IF (583 байт < Количество данных ≤ 157746380 байт) AND (Сеть получателя = 193.125.180.0/24) AND (1 < Порт получателя ≤ 26996) AND (1 < Порт отправителя ≤ 27394) THEN («Чистый» период)

В полученных закономерностях достаточно ясно прослеживается корреляция между аномальной активностью и большими объемами ICMP- и UDP-трафика, причем подозрение вызывают сети 193.124.35.0/24 и 84.237.0.0/17, с активностью которых, связана оценка ситуации системой Discovery как аномальной.

2. *Предсказание на контрольной выборке.* Данный тест использует возможности системы Discovery по построению прогнозов. Перед анализом данных таблицы из нее была удалена 1 000 записей, выбранных произвольным образом. Эти записи были помещены в отдельную таблицу, называемую контрольной выборкой. После успешного обучения системы Discovery на основных данных, алгоритм обрабатывает контрольную выборку, пытаясь восстановить значение признака «Anomalous» по набору значений остальных признаков, на основе найденный закономерностей, т. е. алгоритм определяет, какому периоду, «чистому» или нет, принадлежат записи контрольной таблицы. Результаты теста показали, что используемая реализация системы Discovery с вероятностью около 71 % верно определяет значение целевого признака.

Стоит отметить, что этот результат лучше, чем результат, который дают стандартные алгоритмы Data Mining, основанные на поиске закономерностей в виде логических формул и применяемые Microsoft SQL Server 2005. Так, например, алгоритм Microsoft Association Rules, также извлекающий закономерности из данных в виде правил, показал точность около 64 %.

3. *Оценка количества нарушенных закономерностей на контрольной выборке.* В этом тесте система Discovery не использовала на этапе обучения информации об аномальной ситуации. Целью обучения являлось обнаружение высоковероятностных закономерностей, описывающих нормальное состояние системы. Затем была осуществлена проверка выполнения этих закономерностей на контрольной выборке. Контрольная выборка состояла из 1 000 записей, выбранных произвольным образом как из данных «чистого» периода, так и из данных периода, в котором наблюдались аномалии. Перед обучением данные контрольной выборки, соответствующие «чистому» периоду, были удалены из обучающей таблицы.

На этот раз в правилах фигурировали только шесть признаков, описанных ранее, так как значение признака «Anomalous» тождественно равнялось нулю. В качестве целевого признака использовались все шесть признаков. Пороговые значения точного критерия Фишера и критерия Юла равнялись соответственно 0,05 и 0,5. Для проверки закономерностей на контрольной выборке использовались только правила, УЧ которых больше 0,9.

В результате на одной записи контрольной таблицы, соответствующей «чистому» периоду, нарушается в среднем 0,47 закономерности. На одной записи контрольной таблицы, соответствующей аномальному периоду, нарушается в среднем 3,37 закономерности.

Для контрольной выборки были оценены ошибки первого и второго рода. Для ошибки первого рода было подобрано пороговое значение, при котором количество записей кон-

трольной таблицы за «чистый» период с уровнем аномальности больше порогового было бы меньше 3 % от общего количества записей за «чистый» период. Уровень аномальности записи определяется следующим образом:

$$\sum_{i=1}^{N} -\ln(1-\text{УЧ}_i),$$

где N – число нарушившихся правил на данной записи таблицы, УЧ_i – условная вероятность i-го правила.

В эксперименте пороговое значение было автоматически вычислено, чтобы обеспечить ошибку первого рода не более 3 %. Оно получилось равным 6,5. При таком пороговом значении для записей таблицы, на которых наблюдается превышение установленного порога, нарушается в среднем 5–12 закономерностей на нормальных записях и 10–25 закономерностей на аномальных.

Список литературы

1. *Weiss G.* Mining with Rarity: a Unifying Framework // ACM SIGKDD Explorations Newsletter. 2004. Vol. 6. Issue 1.

2. *Lin S.*, *Chalupsky H.* Unsupervised Link Discovery in Multi-relational Data via Rarity Analysis // Proc. the 3rd IEEE International Conference on Data Mining, ICDM '03. Melbourne, Florida, USA, 2003.

3. *Rattigan M.*, *Jensen D.* The Case for Anomalous Link Detection // Proc. of the Fourth International Workshop on Multi-Relational Data Mining (MRDM-2005). Chicago, 2005. (http://kdl.cs.umass.edu/papers/rattigan-jensen-mrdm2005.pdf)

4. *Getoor L.* Link Mining: A New Data Mining Challenge // SIGKDD Explorations. 2003. Vol. 5. Is. 1.

5. *Badia A.*, *Kantardzic M.* Link Analysis Tools for Intelligence and Counterterrorism // Intelligence and Security Informatics: Proc. of IEEE International Conference on Intelligence and Security Informatics, ISI 2005. Atlanta, GA, USA, 2005.

6. *Kovalerchuk B.*, *Vityaev E.* Data Mining in Finance: Advances in Relational and Hybrid Methods. Kluwer, 2000. 308 p.

7. *Vityaev E.*, *Kovalerchuk B.* Empirical Theories Discovery Based on the Measurement Theory // Mind and Machine. 2004. Vol. 14. No. 4. P. 551–573.

8. *Vityaev E.*, *Kovalerchuk B.* Visual Data Mining with Simultaneous Rescaling // Visual and Spatial Analysis. Advances in Data Mining, Reasoning and Problem Solving. Springer, 2004. P. 371–385.

9. *Витяев Е. Е.* Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов: Моногр. / Новосиб. гос. ун-т. Новосибирск, 2006. 293 с.

10. *Kovalerchuk B.*, *Vityaev E.* Correlation of Complex Evidences and Link Discovery. The Fifth International Conference on Forensic Statistics (Venice International University, Isola di San Servolo, Venice, Italy, Aug. 30 – Sept. 2, 2002), Venice, 2002.

11. *Kovalerchuk B.*, *Vityaev E.* Detecting Patterns of Fraudulent Behavior in Forensic Accounting // Proc. of the Seventh International Conference «Knowledge-based Intelligent Information and Engineering on Systems», Lecture Notes in Computer Science. V. 2773/2003, Oxford, UK, 2003. Vol. 2773/2003, part 1. P. 502–509.

12. *Kovalerchuk B.*, *Vityaev E.*, *Ruiz J. F.* Consistent and Complete Data and «Expert» Mining in Medicine // Medical Data Mining and Knowledge Discovery. Springer, 2001. P. 238–280.

13. *Kovalerchuk B.*, *Vityaev E.*, *Ruiz J. F.* Consistent Knowledge Discovery in Medical Diagnosis // IEEE Engineering in Medicine and Biology Magazine (Special issue on Data Mining and Knowledge Discovery). 2000. Vol. 19. No. 4. P. 26–37.

14. *Витяев Е. Е.* Семантический подход к созданию баз знаний. Семантический вероятностный вывод наилучших для предсказания ПРОЛОГ-программ по вероятностной модели данных // Логика и семантическое программирование. Новосибирск, 1992. Вып. 146. С. 19–49.

15. *Витяев Е. Е.*, *Москвитин А. А.* Введение в теорию открытий. Программная система DISCOVERY // Логические методы в информатике. Новосибирск, 1993. Вып. 148. С. 117–163.

16. *Кендал М.*, *Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973.

17. *Halpern J. Y.* An Analysis of First-Order Logic of Probability // Artificial Intelligence. 1990. Vol. 46. P. 311–350.

Материал поступил в редколлегию 17.06.2008

E. E. Vityaev, B. K. Kovalerchuk, A. M. Fedotov,
D. S. Durdin, A. V. Demin, S. D. Belov, V. B. Barakhnin

Regularities Discovery and Pattern Recognition of the Abnormal Events in the Network Traffic Data Stream

A new approach to the abnormal events detection is presented in the paper based on the hybrid events correlation (HEC), developed by the authors. The main idea of the HEC is detection of the abnormal events as infraction of the normal events process. The normal events process is determined by the set of high probability regularities that are discovered on the set of normal events process data and determine the laws of the normal events process. The high probability regularities were discovered by the Discovery system, that realizes the developed by authors relational approach to the intelligent data analysis. The system was applied to the analysis of the server traffic of the SB RAS data transferring system. The significant difference in regularities of the normal and abnormal events process was detected.

Keywords: Data Mining and Knowledge Discovery in Data Bases, Link Discovery, Fraud detection, abnormal events.