Шудрак Максим Олегович – Сибирский государственный аэрокосмический университет им. М.Ф. Решетнева (СибГАУ); e-mail: mxmssh@gmail.com; 660014, г. Красноярск, пр. им. газеты «Красноярский рабочий», 31; тел.: 89233088703; кафедра безопасности информационных технологий; аспирант.

Лубкин Иван Александрович – e-mail: lubkin@rambler.ru; тел.: 83912328627; кафедра безопасности информационных технологий; аспирант.

Золотарев Вячеслав Владимирович – e-mail: amida@land.ru; тел.: 89050874847; кафедра безопасности информационных технологий; к.т.н.; доцент.

Shudrak Maxim Olegovich – Siberian Aerospace State University named after M.F. Reshetnev (SibSAU); e-mail: mxmssh@gmail.com; 31, Krasnoyarsky Rabochy av., Krasnoyarsk, 660014, Russia; phone: +79233088703; the department of information technology security; postgraduate student.

Lubkin Ivan Alexandrovich – e-mail: lubkin@rambler.ru; phone: +73912328627; the department of information technology security; postgraduate student.

Zolotarev Vyacheslav Vladimirovich – e-mail: amida@land.ru; phone: +79050874847; the department of information technology security; cand. of eng. sc.; associated professor.

УДК 004.056

А.А. Таран

ПРИЛОЖЕНИЯ АЛГОРИТМА ANTMINER+ К ЗАДАЧЕ КЛАССИФИКАЦИИ СОБЫТИЙ ПРИ АНАЛИЗЕ СЕТЕВОГО ТРАФИКА

Статья посвящена исследованию технологий классификации данных с помощью алгоритма AntMiner+ и в частности их приложениям к обнаружению вторжений при исследовании сетевого трафика. Особое внимание при этом уделено свойствам алгоритма, позволяющим автоматически получать понятные, читаемые и явно связанные с предметной областью задачи описания исследуемых множеств. Сделан вывод о применимости рассматриваемого алгоритма к задачам автоматической генерации сигнатур атак и профилей нормального поведения системы. Описаны результаты экспериментов, подтверждающие выдвинутые гипотезы о свойствах алгоритма.

Обнаружение вторжений; классификация данных; анализ сетевого трафика; AntMiner+; аномалии; сигнатуры; извлечение правил.

A.A. Taran

APPLICATION OF ANTMINER+ ALGORITHM TO EVENT CLASSIFICATION FOR NETWORK TRAFFIC ANALYSIS

The article deals with the technologies of data classification based on the algorithm AntMiner+ in the analysis of the network traffic for intrusion detection. Special attention is paid to the properties of the algorithm which allow us automatically receive understandable, human readable and strongly related with the problem under consideration and describing sets. Some conclusion about applicability of the methods to the tasks of automated generation of attack's signatures and profiles of system's normal behavior are made. The paper also explores the results of experiments which confirm the hypothesis about algorithm's properties.

Anomaly detection; data classification; network traffic analysis; AntMiner+; signatures; rule conduction.

Введение. Системы обнаружения вторжений (СОВ) наряду с криптографическими средствами, как правило, составляют основу организации противодействия угрозам несанкционированного доступа к данным в компьютерной системе, осо-

бенно при наличии подключения к локальной сети или сети Интернет. При этом СОВ предоставляют более гибкое решение, так как, например не требуют реализации фиксированных протоколов передачи данных для всех возможных участников взаимодействия с системой. Кроме того при реализации систем обнаружения вторжений можно подойти к проблеме более комплексно, анализируя не только данные проникающие из вне, но и внутренние процессы (интенсивность работы с файловой системой, потоки данных во время исполнения программ и т.д.). Вот почему реализация СОВ, позволяющих обнаруживать вторжения на ранних этапах с минимальными затратами ресурсов и незначительной вероятностью появления ошибок первого и второго рода представляется актуальной. Решение этой задачи требует построения специальной алгоритмической и математической базы для эффективного анализа и классификации данных.

Принцип работы СОВ можно описать следующим образом. В защищаемой компьютерной системе или сети выделяется пространство анализируемых событий, специфичных для данной системы. Каждому из событий в этом пространстве ставится в соответствие набор значений характеристик, определенных заранее. Их вид и количество выбирается таким образом, чтобы данный набор позволял однозначно отнести любое событие к множеству нормальных событий или к отклонениям от этой нормы. Разбиение всех событий на эти два множества как раз и является основной задачей СОВ. При этом обнаруженные аномалии система считает вторжениями. В такой постановке естественно применять методы, разработанные для решения классических задач интеллектуального анализа данных: классификации и кластерного анализа.

В данной работе рассмотрена возможность применения алгоритма AntMiner+, относящегося к недавно получившей популярность области интеллектуального анализа данных – системам роевого интеллекта. В основе данного направления лежит заимствованная из биологии идея подражания естественным самоорганизующимся децентрализованным системам, таким как, например, пчелиный рой, птичья стая, муравьиная колония и т.д.

В последнем случае речь идет о технологии Ant Colony Optimization (муравьином алгоритме). Она основана на аналогиях с поведением колонии муравьев, занимающихся поиском пищи, которые используют след из испаряющихся и вновь усиливаемых феромонов для определения кратчайшего пути к источнику питания. В литературе можно найти применения методов Ant Colony Optimization к анализу данных и даже в задачах выявления вторжений (например, в [1]). В частности используются значения феромона, их испарение и усиление, переходные вероятности, а также множество агентов, производящих операции с данными. Однако в этом случае авторы обращаются к модификациям классических методов кластерного анализа и не используют некоторые особые свойства, имеющиеся непосредственно у AntMiner+. В данной работе сделана попытка обратиться именно к этим специфическим особенностям для получения максимального эффекта от классификации событий в задачах обнаружения вторжений. В частности рассматривается генерация явно связанных с предметной областью результатов для получения легко читаемых человеком сигнатур атак или профилей нормального поведения системы.

Алгоритм AntMiner+. Муравьиные алгоритмы вообще и AntMiner+ в частности являются итерационными, получая оптимальное решение, последовательно улучшая построенное ранее с помощью множества агентов. Непосредственно AntMiner+ может быть применен к решению задач, в которых конечный результат может быть представлен в виде набора составных частей определенного вида. С каждой такой частью ассоциируется вершина графа. Его ребра соединяют вершины, соответствующие частям, которые могут одновременно встретиться в кон-

кретном решении. Так, например, при оценке характеристик нормальных событий в компьютерной системе вершины (узлы) рассматриваемого графа связываются с ограничениями, накладываемыми на отдельные характеристики. При последующем анализе данных будут рассматриваться три вида ограничений: равенство значения характеристики фиксированному значению, принадлежность некоторому интервалу значений и отсутствие ограничений. Будем обозначать $V_{i,j}$ вершину, сопоставляющую i-ой характеристике j-ое ограничение. В каждом таком графе (рис. 1) выделяют два особых узла: начальный (start) и финальный (finish).

В качестве исходных данных на вход алгоритма подаются два множества векторов заданной размерности, в результате обработки этих данных с помощью рассматриваемого метода должен быть сконструирован набор правил, позволяющий выбрать как можно больше точек из первого (целевого множества) и как можно меньше из второго. Фактически, цель алгоритма – построить ограничения, выделяющие из непрерывного потока смешанных данных элементы, принадлежащие только целевому множеству. Физический смысл каждого из двух наборов векторов зависит от контекста задачи и, как будет показано далее, может меняться в зависимости от целей аналитика, применяющего алгоритм.

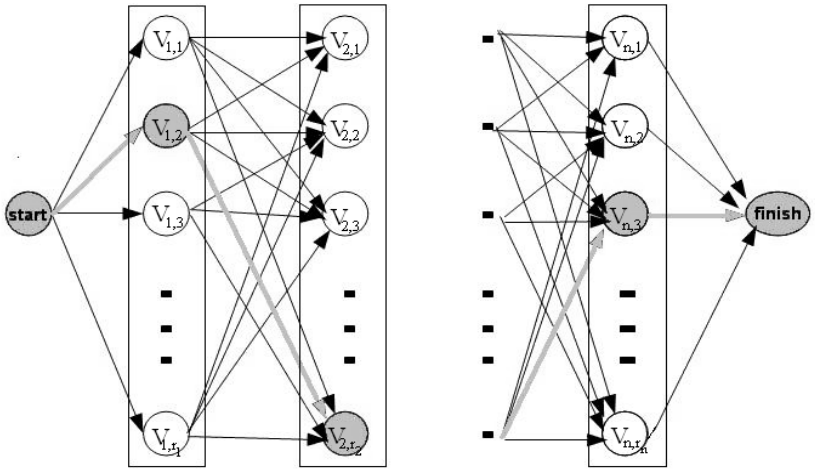

Рис. 1. Структура базового графа метода

Для связи построенного графа с входными данными для каждого узла V_{ij} рассчитывается значение эвристики η_{ij}, показывающее насколько выбранная часть решения соответствует возможному оптимальному результату. В данной работе значение η_{ij} вычисляется как отношение числа точек, принадлежащих целевому классу, у которых в то же время значение i-ой характеристики удовлетворяет j-му ограничению, к числу точек, для которых выполняется только вторая часть утверждения. Общее качество полученного решения определяется с помощью функции вещественного переменного δ, заданной следующим образом:

$$\delta = \frac{TP}{TP + FP} + \frac{TP}{All},$$

где *All* – общее число точек, содержащихся в обоих множествах, *TP* – число точек из целевого множества, охваченных правилом, а *FP* – число точек, охваченных правилом, но не принадлежащих целевому множеству. Таким образом, максимизируя функцию δ, вероятнее всего можно получить оптимальный набор правил.

Важную роль в работе алгоритма играет параметр τ, так называемый уровень феромона. Каждому ребру в построенном графе сопоставляется свое значение этого параметра, которое изменяется в конце каждой итерации алгоритма двумя способами: испарение и усиление. Испарение – уменьшение величины параметра умножением на заданный коэффициент ρ (|ρ|<1). Эта операция производится над всеми ребрами описанного графа. Усиление же влияет только на ребра, связывающие части лучшего решения для данной итерации. Это действие состоит в увеличении τ на значение функции δ – оценку качества лучшего решения. Исходя из параметров τ и η, для каждого узла рассчитываются вероятности перехода в любой другой узел. Заданное количество муравьев-агентов путешествует по графу в направлении от начальной вершины к финальной в соответствии с этими вероятностями. Каждый агент запоминает пройденный путь, который и становится решением-кандидатом. После этого производится оценка выработанного маршрута (правил, соответствующих его узлам), а также пересчет уровней феромона и переходных вероятностей. На этом итерация завершается. Данный процесс повторяется, пока не будет достигнуто условие остановки. В описанных ниже экспериментах таким условием считается совпадение наилучших сгенерированных правил в течение нескольких итераций. После остановки итерационного процесса лучшее правило попадает в результирующий набор. Количество сгенерированных правил ограничивается с помощью фиксированной величины покрытия правилами точек целевого множества. В рассмотренных далее примерах использована граница 96 %. Более подробно с описанным выше методом можно ознакомиться в [2].

Автоматическая генерация сигнатур. Одна из интересных для задачи классификации особенностей метода AntMiner+ – форма представления результата. В конечном итоге при удачном завершении работы алгоритма конструируется набор ограничений для анализируемых характеристик, определяющих принадлежность точек, удовлетворяющих этим ограничениям, целевому множеству. При этом все значения характеристик рассматриваются без усреднения, в их естественном виде, а значит, полученные правила позволяют описывать структуру целевого множества в понятной человеку форме. Кроме того, поскольку для реализации метода не требуется проводить вычисления непосредственно над входными данными, то типы рассматриваемых характеристик могут быть не только числовыми, но и качественными. При этом не требуется вводить дополнительные метрики для работы с такими параметрами. А значит, в процессе анализа в данных не проявятся искусственные закономерности, которых нет в реальной предметной области. Все эти рассуждения приводят к выводу о возможной эффективности применения AntMiner+ к задаче автоматической генерации сигнатур атак.

В частности данный подход подход был применён для анализа данных в рамках двух экспериментов. Первый – анализ icmp-соединений из сборника, созданного для соревнования KDD CUP'99 [3], с целью выявления значений характеристик, указывающих на наличие smurf-атаки (ее сигнатуры).

Smurf – вид атаки отказ в обслуживании, заключающийся в рассылке широковещательных ping-запросов с использованием в качестве адреса источника пакетов IP адреса жертвы. Для поиска сигнатуры из множества всех описанных в KDD CUP параметров было выбрано, 10 характеристик, часть из которых имеет символьный тип, другая – вещественный. В описании к условиям соревнований они помещены под следующими именами: *service, src_bytes, count, srv_ count, dst_host_count, dst_host_srv_count, dst_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_port_rate*, *dst_host_srv_diff_host_rate*. Вектора, состоящие из значений этих характеристик для соединений, помеченных как smurf-атака, поданы на вход алгоритма в качестве целевого множества. Второе множество сформировано

из векторов значений выбранных характеристик для всех остальных icmp-соединений. В результаты работы метода для описанных входных данных сгенерировано 2 правила:

$$dst\_host\_same\_port\_rate == 1$$

$$510 \le srv\_count \le 511$$

Первое выбирает из обоих множеств все соединения, у которых предыдущие коммуникации между источником и приемником приходились на один порт. Второе правило позволяет выбрать из оставшихся векторов те соединения, число взаимодействий между участниками которого за последние две секунды является максимальным среди рассмотренных. Такое описание дает в целом верное, хотя и недостаточно полное, представление о рассмотренной атаке. При применении первого правила из 3402 нормальных соединений 1575 были ошибочно отнесены к smurf-атаке, а 1492 smurf-соединений из 164591 оказались не охваченными. Следующее правило позволило избавиться от ошибок при классификации, однако осталось 432 не классифицированных smurf-соединения. На рис. 2 демонстрирует эффективность полученных правил.

В другом случае с помощью алгоритма AntMiner+ было организовано исследование сетевых соединений, данные о которых получены на одном из серверов факультета в течение суток для получения правила, позволяющего выделить соединения, проходящие через порт 80 от остальных.

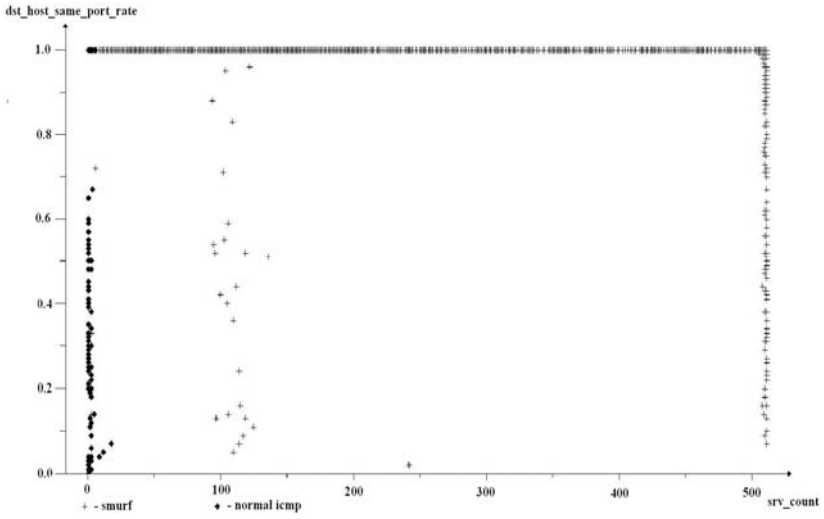

Рис. 2. Проекция для поиска сигнатуры smurf-атаки

Для упрощения описания задачи из всей информации, собранной на сервере, были использованы только данные о соединениях, приходящихся на 80 и 22 порт. С помощью пяти характеристик, а именно: продолжительность соединения (*duration*), число пакетов (*from_packets*) и байт (*from_bytes*) от инициатора соединения и к нему (*to_packets*, *to_bytes*) – была предпринята попытка отделить друг от друга эти два множества. При этом в качестве целевого было выбрано множество соединений, связанных с 80 портом. Сгенерированный набор правил:

$$4 \le from\_packets \le 6$$

$$(7 \le from\_packets \le 173) and (7 \le to\_packets \le 117) and (4312 \le from\_bytes \le 229372)$$

Эти ограничения классифицируют исходные данные со стопроцентной точностью и лишь 4 % точек остаются неохваченными (рис. 3).

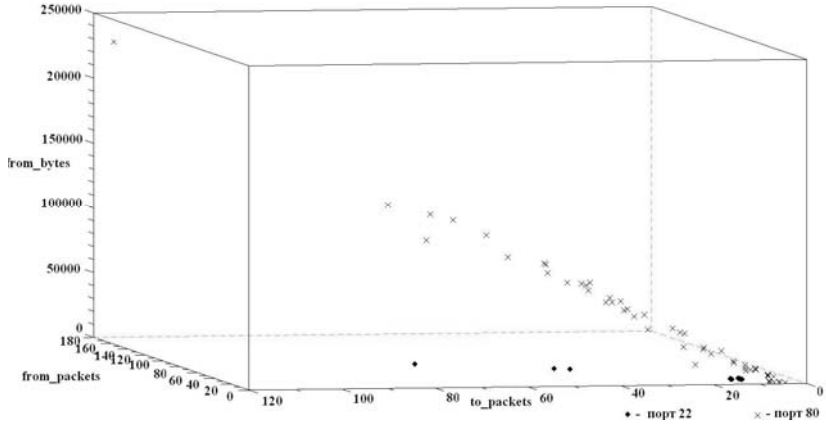

Рис. 3. Результат проектирования на выбранные характеристики при анализе подключений к 80 и 22 портам

Более точные результаты во втором эксперименте вероятнее всего связаны с тем, что в данном случае в отличие от набора KDD CUP'99 рассматриваются однородные хорошо структурированные множества, в которых легко обнаружить необходимые закономерности. Поэтому можно сделать вывод о необходимости применения классических методов кластерного анализа на этапе предшествующем работе алгоритма AntMiner+ для выделения множеств с более четкой организацией, тогда описанный метод позволит легко получить точное описание закономерностей в полученных множествах.

Профиль нормального поведения системы и обнаружение аномалий. Продолжая анализировать задачу разбиения соединений по портам назначения, можно рассмотреть способы поиска аномалий на основе применения AntMiner+. В качестве целевого множества на вход алгоритма передается набор векторов, содержащий смесь данных, соответствующих соединениям с 80 и 22 портом, а другое множество состоит только из соединений, относящихся к 22 порту. Условия выхода из алгоритма - появление правила, не накладывающего ограничений ни на одну из характеристик. В результате получается набор правил, идентичный сгенерированному ранее. Таким образом, применяя описанную технику, можно не только обнаружить аномалию в определенном наборе данных, но и одновременно получить ее описание в понятной, читаемой форме.

Другое важное свойство рассмотренной модификации муравьиного алгоритма, о котором необходимо упомянуть, заключается в том, что в некоторых случаях, объединяя в одно правило ограничения, накладываемые на разные характеристики, AntMiner+ позволяет не только отделять друг от друга элементы разных наборов данных, но и искать связи между характеристиками внутри целевого множества. Это свойство может быть применено, например, при анализе программного кода для поиска аномалий в аргументах системных вызовов. Уже разработано множество алгоритмических и программных моделей, использующих соотношения между аргументами вызовов для повышения эффективности защиты компьютерных систем с помощью хост-ориентированных СОВ (см., например, [4]). С другой стороны эти закономерности могут становиться основой для генерации профилей нормального поведения системы, что является одной из самых мно-

госторонних и сложных задач при решении проблемы обнаружения вторжений. Поэтому возможность автоматизированного поиска подобных соотношений очень ценна для разработчиков систем защиты.

Для получения исходных данных при реализации эксперимента, подтверждающем данную гипотезу, была написана программа на языке C, совершающая в цикле следующую последовательность действий: открытие случайного файла из текущего каталога для чтения, чтение его первого байта, закрытие файла, открытие файла “info.txt” для записи, запись прочитанного байта в файл, закрытие файла. В качестве целевого множества используется набор всех использованных в данной программе аргументов системного вызова open, а именно: имя файла

$$\text{mode} = \text{O\_RDONLY};$$

$$(\text{name} = \text{"/home/user/info.txt"}) \text{and} (\text{mode} = \text{O\_WRONLY})$$

(name) и режим открытия (mode). Во входное множество алгоритма AntMiner+, не являющееся целевым, помещаются значения аргументов, не встречающиеся в программе. В данном случае – это name = “” , mode = O_RDWR, поскольку в программе используются режимы O_RDONLY и O_WRONLY, а файлов с таким именем не существует. Количество точек в обоих множествах должно быть примерно одинаковым. В результате получен следующий набор правил:

Таким образом, условия чтения и записи файлов корректно определены, а значит выявление закономерностей среди значений различных характеристик с помощью алгоритма AntMiner+ действительно возможно.

Предварительная обработка и снижение размерности данных. Наравне с рассмотренными достоинствами алгоритма AntMiner+ следует отметить, что время работы метода в значительной степени зависит от количества анализируемых характеристик, среднего количества возможных ограничений для конкретной характеристики и общего числа обрабатываемых данных. Пусть *layers* – количество характеристик, а *avg* – среднее число значений для каждой, тогда число вершин графа оценивается значением $layers \cdot avg$, а количество ребер, для которых нужно производить расчеты, величиной $avg + (layers - 1) \cdot avg^2$. Причем для вершин требуется получить значения эвристики, а для ребер — переходные вероятности на каждой итерации алгоритма. Кроме того, пересчет значений эвристики η, как и качества решений δ требует анализировать все элементы входных множеств. Таким образом, рассматриваемый алгоритм довольно ресурсоемкий и требует предварительной обработки входных данных до начала анализа для повышения эффективности.

В частности, в примере с набором KDD CUP’99 для достижения разумного времени получения результата требуется минимум в 4 раза уменьшить количество предлагаемых характеристик. Для этого, во-первых, значения всех вещественных характеристик разбиваются на интервалы, содержащие примерно одинаковое количество точек. Эту операцию надо проводить особенно аккуратно, поскольку при попадании в интервал большого числа значений, не принадлежащих целевому множеству, система не сможет построить верное решение. Далее из множества векторов, соответствующих icmp-соединениям выбираются только те характеристики, которые принимают более одного значения, причем для любого фиксированного значения этих характеристик найдется достаточно большое количество элементов, отличных от него. Конечно, в этом случае можно использовать известные алгоритмы редукции размерности, такие например, как метод главных компонент или линейный дискриминантный анализ (см. [5]). Но недостатком этих алгоритмов является их возможность работать только с числовыми данными, кроме того, исходные значения подвергаются преобразованиям (центрирование, нор-

мировка), которые меняют структуру исследуемых множеств, а значит, полученные впоследствии результаты будут в форме, неудобной для понимания человеком или соотнесения с реальной предметной областью, что потребует дополнительных временных затрат от аналитика.

Интересно, что сам муравьиный алгоритм в некоторых случаях также может быть полезен для решения задач снижения размерности. Поскольку характеристики, не включенные в сгенерированные правила, признаются методом бесполезными для классификации, а значит их можно учитывать в последнюю очередь при анализе тех же данных другими способами.

Выводы. Опираясь на описанные выше свойства алгоритма и результаты экспериментов, можно говорить об эффективном применении алгоритма AntMiner+ в качестве вспомогательного инструмента аналитика по всем направлениям решения задачи обнаружения вторжений (генерация сигнатур и профилей нормального поведения, поиск аномалий и снижение размерности данных). Тем не менее, данный подход неприменим к системам реального времени и в задачах, требующих высокой точности результатов в связи с низкой скоростью обработки информации и ощутимой зависимостью генерируемых правил от структуры начальных данных. Поэтому необходимы исследования возможности модификации алгоритма, в частности оптимизации расчета эвристики, вероятностей и качества решений, а также улучшения максимизируемой функции качества для получения более точных результатов. Однако уже сейчас описанный метод может служить эффективным инструментом для специалистов, изучающих наборы данных любого вида с целью выявления структуры и закономерностей среди характеристик исследуемых множеств (в частности в рамках задачи обнаружения вторжений). По этой причине алгоритм также является хорошим дополнением к классическим методам реализации СОВ.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Tsang C.-H., Kwong S.* Ant colony clustering and feature extraction for anomaly intrusion detection. 2006. Studies in computational intelligence. – Vol. 34. Springer. – P. 101-121.
2. *Martens D., De Backer M., Haesen R., Baesens B., Holvoet T.* Ants constructing rule-based classifiers. 2006. Studies in computational intelligence, vol. 34. Springer. – P. 21-41 .
3. Официальный сайт KDD CUP. http://www.sigkdd.org/kddcup/index.php? section=1999&method=task.
4. *Bhatkar S., Chaturvedi A., Sekar R.* Dataflow Anomaly Detection // SP '06 Proceedings of the 2006 IEEE Symposium on Security and Privacy. IEEE Computer Society Washington, DC, USA. 2002. – P. 48-62.
5. *Нестеренко В. А., Таран А.А.*, Редукция размерности пространства состояний в задачах анализа сетевого трафика // Известия ЮФУ. Технические науки. – 2011. – № 12 (125). – С. 96-103.

Статью рекомендовал к опубликованию д.т.н. профессор Р.А. Нейдорф.

Таран Анна Александровна – Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Южный федеральный университет»; e-mail: annie4ka@yandex.ru; 34413, г. Ростов-на-Дону, ул. Добровольского, 36/2, кв. 115; тел.: 88632749704, 89515034220; кафедра информатики и вычислительного эксперимента, бакалавр

Taran Anna Alexandrovna – Federal State-Owned Autonomy Educational Establishment of Higher Vocational Education “Southern Federal University”; e-mail: annie4ka@yandex.ru; fl. 115, 36/2, Dobrovolskogo street, Rostov-on-Don, 344113, Russia; phone: +79515034220, +78632749704; the department of computer science and computer simulation; bachelor.