УДК 004.056.53

Попов Е. Ф., Тюкова А. А., Фучко М. М., Захаров А. А.

ВЫЯВЛЕНИЕ НЕТИПИЧНЫХ СОБЫТИЙ СРЕДСТВАМИ СТАТИСТИЧЕСКОГО АНАЛИЗА

В работе изучается вопрос о применимости алгоритмов статистического анализа для выявления нетипичных событий и состояний в различных информационных системах. Рассматривается возможность применения кластерного анализа для выявления нетипичных потоков трафика в сети передачи данных, а так же его эффективность при различной интерпретации характеристик потоков сетевого трафика.

Ключевые слова. *Статистический анализ, кластерный анализ, сети передачи данных, выявление аномалий, система обнаружения вторжений.*

Попов Е. Ф., Тюкова А. А., Фучко М. М., Захаров А. А.

ВЫЯВЛЕНИЕ НЕТИПИЧНЫХ СОБЫТИЙ СРЕДСТВАМИ СТАТИСТИЧЕСКОГО АНАЛИЗА

В работе изучается вопрос о применимости алгоритмов статистического анализа для выявления нетипичных событий и состояний в различных информационных системах. Рассматривается возможность применения кластерного анализа для выявления нетипичных потоков трафика в сети передачи данных, а так же его эффективность при различной интерпретации характеристик потоков сетевого трафика.

Ключевые слова. *Статистический анализ, кластерный анализ, сети передачи данных, выявление аномалий, система обнаружения вторжений.*

Введение

В условиях быстрого развития информационных технологий постоянно разрабатываются новые методы атак на информационные системы, что требует непрерывного совершенствования средств защиты информации. Так как предусмотреть все способы атак на информационную систему зачастую не представляется возможным, оптимальным решением является выявление нетипичных событий и состояний системы.

Для выявления отклонений необходимы показатели, которые отображают типичное состояние информационной системы и могут выступать в качестве эталона для сопоставления с данными, обрабатываемыми в реальном времени. Применение алгоритмов статистического анализа к данным, собранным в период штатного функционирования, позволяет выделить статистические показатели, при сравнении с которыми можно выявить нетипичные события и состояния, являющиеся потенциально опасными для информационной системы.

В данной работе изучается вопрос о применимости алгоритмов статистического анализа, для выявления нетипичных событий и состояний на примере потоков трафика в се-

тях передачи данных. Рассматривается эффективность применения результатов кластерного анализа в реальном времени при различной интерпретации характеристик потоков сетевого трафика. По результатам применения кластерного анализа к потокам трафика в сети передачи данных оценивается возможность применения используемого метода для выявления нетипичных событий и состояний в различных подсистемах и элементах информационной инфраструктуры.

Сбор и интерпретация статистических данных

Сложность применения статистического анализа в данном контексте заключается в том, что применяемые алгоритмы должны быть адаптированы для обработки данных в реальном времени и должны производить интеллектуальный анализ, учитывающий не только текущее состояние системы или состояния за короткий промежуток времени. Наиболее ценными являются данные, собранные за длительные периоды штатного функционирования, которые являются необходимыми для наиболее точного выявления типичных событий и состояний системы.[1]

Для обработки характеристик, собранных за длительный период, важно применять алгоритмы статистического анализа, не только адаптированные для обработки большого количества данных, но и способные учитывать степень устаревания, что является необходимым для адаптации к изменениям в информационной системе. Наиболее оптимальными являются алгоритмы, которые способны производить анализ не только на базе набора статистических данных, но и с учетом предыдущих результатов работы алгоритмов статистического анализа.

События и состояния могут рассматриваться как объекты, обладающие рядом характеристик. Полученные объекты могут быть использованы в качестве статистических единиц для таких алгоритмов статистического анализа, как алгоритмы кластеризации. Рассмотрим применение статистического анализа для выявления нетипичных состояний на примере.

Наиболее удобным для рассмотрения является пример применения статистического анализа к потокам трафика в сети передачи данных. Для сбора информации о потоках трафика могут использоваться сенсоры, базирующиеся на протоколе NetFlow. Подобные сенсоры позволяют получить следующие характеристики:

- IP-адреса отправителя и назначения;
- протоколы сетевого и транспортного уровня;
- номера портов (TCP/UDP), позволяющие определить используемый протокол прикладного уровня;
- время;
- объем переданных данных;
- средний размер пакетов данных.[2]

Полученные характеристики необходимо адаптировать для применения кластерного анализа.

IP-адрес не может применяться для кластерного анализа в чистом виде, так как не характеризует никаких особенностей потока трафика, и данные, передаваемые на различные адреса, могут быть предназначены для работы с одним и тем же сервисом. Например, таким как поисковик Google, который обеспечивает балансировку нагрузки за счет соответствия своему основному доменному имени ряда различных IP-адресов, что не всегда возможно отслеживать автоматически в процессе кластерного анализа.[3]

Но за счет адреса назначения можно выявить ряд особенностей потока трафика, например, выявить для внутренней или для внешней сети предназначен поток трафика, что является важной характеристикой в процессе анализа. Так же, опираясь на адрес отправителя, можно выявить какие устройства или какая группа пользователей располагаются в данной подсети, определив из какой подсети инициирован поток трафика.

Протокол прикладного уровня, который определяется номерами портов, является одной из ключевых характеристик, которая может быть интерпретирована различным образом. Можно разделять статистические единицы на множество классов, считая потоки данных каждого из протоколов прикладного уровня отдельными классами. Так же, протоколы прикладного уровня могут быть разделены на классы по их функциональному предназначению:

- получение информации с веб-сайтов;
- передача файлов;
- мгновенный обмен сообщениями;
- потоковая передача данных;
- удаленное управление;
- и т.д.

При сравнении потоков сетевого трафика, относящихся к различным классам, было

выявлено, что остальные характеристики, такие как объем передаваемых данных, длительность передачи данных, средний размер пакета зависит от класса потока трафика, определенного по функциональному предназначению протокола прикладного уровня. И характеристики потоков сетевого трафика различных протоколов прикладного уровня, относящихся к одному классу, имеют сходство намного большее, чем в случае с протоколами, относящимся к разным классам.[4]

Применение статистического анализа

Необходимо выделить единицу потока трафика, к которой будет применяться статистический анализ. В данном контексте статистической единицей могут являться:

- поток трафика от уникального отправителя уникальному получателю за весь период передачи данных (период активности сессии);
- поток трафика от уникального отправителя уникальному получателю за заданный промежуток времени.

Анализ потока за полный период передачи данных между двумя узлами предоставляет наиболее точную характеристику природы трафика и является наиболее ценным для статистического анализа. Но статистика, составленная на базе информации о сетевом трафике, переданном за полный период активности сессии, может быть использована для обнаружения нетипичных потоков данных, только после завершения сессии, что значительно увеличивает время реакции системы.

При использовании в качестве статистических единиц сегментов, выделенных из общего потока, ограниченных небольшими промежутками времени, например интервалами в 30 секунд, снижается точность статистического анализа, но появляется возможность использования результата статистического анализа для обнаружения нетипичных действий в реальном времени, с максимальным временем реакции равным выбранному интервалу времени для статистической единицы.[5]

Каждую отдельную статистическую единицу можно представить в виде точки, расположенной в многомерном пространстве, каждое из измерений которого представляет собой одну из характеристик потока трафика. Применение алгоритма кластеризации позволит выявить области с высокой концентрацией точек и объединить их в кластеры. В результате работы алгоритмы будут получены кластеры, отображающие наборы характеристик, свойственные потокам сетевого трафика, передаваемого при штатном функционировании системы.[6]

На базе результатов вычислений могут быть выделены правила, описывающие значения ряда характеристик, свойственные для полученных кластеров, которые могут стать эталонными для определения типичности потоков трафика. Наличие правил, описывающих типичные потоки трафика, открывает возможность производить в реальном времени контроль, лежат ли значения характеристик текущих потоков трафика в рамках типичных для данной информационной системы.

Выводы

В результате исследования было выявлено, что интерпретация статистических единиц, как объектов обладающих рядом характеристик, позволяет адаптировать статистические данные о событиях и состояниях различных подсистем для анализа с применением алгоритмов кластеризации. Необходимым и достаточным для применения кластерного анализа является наличие исчисляемых характеристик событий или состояний. На эффективность алгоритмов кластеризации влияет равномерность распределения и зависимость характеристик статистических единиц.

Было определено, что по результатам кластерного анализа может быть разработан ряд правил, определяющий типичные значения характеристик статистической единицы, которые могут быть применены для выявления нетипичных событий или состояний в реальном времени без необходимости повторного применения алгоритмов статистического анализа.

На примере применения кластерного анализа для выявления нетипичных потоков трафика в сети передачи данных обнаружена зависимость эффективности алгоритмов кластеризации от интерпретации характеристик статистических единиц. Определено, что классификация статистических единиц по характеристикам, не имеющим числового значения, может оказывать высокое влияние на эффективность кластерного анализа, при условии зависимости от них значений исчисляемых характеристик.

Примечания

1. Babenko G. V., Belov S. V. Identification of network abnormalities using methods of statistical analysis //European researcher – 2011. – Т. 1. – №. 5.

2. Claise D. Cisco Systems NetFlow Services Export Version 9 // IETF RFC 3954 URL: http://www.ietf.org/rfc/rfc3954 (датаобращения 08.12.2014)

3. Barroso L. A., Dean J., Holzle U. Web search for a planet: The Google cluster architecture //Micro, Ieee. – 2003. – Т. 23. – №. 2. – С. 22-28.

4. Soysal M., Schmidt E. G. Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison //Performance Evaluation. – 2010. – Т. 67. – №. 6. – С. 451-467.

5. Костенко С. А. Технология применения многомерного шкалирования и кластерного анализа // Фундаментальные исследования. – 2012. – №. 11. – С. 927-930.

6. Дюран Б., Оделл П. Кластерный анализ //М.: Статистика. – 1977. – Т. 15.

Попов Евгений Фёдорович, аспирант ТюмГУ,efpopov@gmail.com

ТюковаАлександра Александровна, аспирант ТюмГУ, tyukovaaa@kbinform.ru

ФучкоМихаил Михайлович, аспирант ТюмГУ, mikhalich@russia.ru

Захаров Александр Анатольевич, д.т.н., профессор ТюмГУ

нужны переводы