ПРИЛОЖЕНИЕ НА ДЪРВО НА РЕШЕНИЯТА В СИСТЕМИТЕ ЗА ОТКРИВАНЕ НА НАРУШЕНИЯ

Веселина Жечева
Евгения Николова
Бургаски свободен университет

DECISION TREE APPLICATION TO INTRUSION DETECTION SYSTEMS

Veselina Jecheva
Evgeniya Nikolova
Burgas Free University

Abstract: *The purpose of the intrusion detection systems (IDS) is to reveal any violence of the organizations' security policy – unauthorized access from outsiders, rising privileges of authorized users, violation of the confidentiality and/or integrity of system resources. The present paper presents an examination of the current IDS, based on the anomalies (behavioral analysis), where C4.5 algorithm is applied in a host-based scenario in order to describe the normal user activity, using decision tree. As a second step, a cluster analysis has been applied with purpose to classify current user activity as normal or malicious. With purpose of approving the proposed methodology, a number of simulation experiments have been applied and the obtained results have been analyzed.*

Key words. *Intrusion detection systems (IDS), anomaly-based IDS, C4.5 algorithm, decision tree, cluster analysis*

1. Въведение

Информационната сигурност е сред критично важните елементи на функционирането на съвременните информационни системи, работещи в мрежова и Интернет среда. Компрометирането на информационната сигурност може да доведе до разкриване, изменение или изтриване на критични за организацията данни, което причинява сериозни загуби.

За решаване на този проблем се прилагат редица утвърдени в практиката решения. Част от тях поддържат нормалната работа на системата (антивирусни програми, защитни стени, системи, предпазващи от рекламен и шпионски софтуер, анти-спам филтри и др.), докато други служат за архивиране и възстановяване на нормалната работа, ако нарушението на сигурността все пак стане факт, както и за откриване и неутрализиране на случайни, но достатъчно сериозни заплахи, които е трудно предварително да бъдат планирани и открити.

От последния вид важно значение за осигуряване работата на големи системи с множество потребители и високи изисквания по отношение на информационната сигурност, имат системите за откриване на нарушения (intrusion detection systems,

IDS). Те са предназначени за откриване на нарушения на политиката на сигурност на дадена организация – неоторизиран достъп на външни лица, повишаване на правата за достъп на легитимни потребители, нарушаване на поверителността и/или цялостността на системни ресурси, използване на чужд профил и др.. За тази цел те съхраняват данни за наблюдаваните събития, уведомяват системните администратори за важни събития и генерират отчети за следените събития [1].

Относно местоположението и обработваните данни тези системи се разделят на хост-базирани (обработват данни от даден хост, т.е. сървър с важни системни услуги), мрежово-базирани (обработват данни от мрежа или мрежов сегмент) и хибридни, съчетаващи и двата метода.

За разпознаване на нарушения се прилагат следните два основни метода: *откриване на сигнатури* и *откриване на аномалии*. Системите, базирани на сигнатури (познати и като базирани на злоупотреби), търсят образци на известни заплахи в данните за текущите събития в системата. Подобно на антивирусните програми, тези системи са ефективни при разпознаване на вече известни атаки, но се справят незадоволително при откриване на нови или варианти на съществуващи атаки. Системите, базирани на аномалии (поведенчески анализ), се основават на идеята, че нарушенията предизвикват отклонение от нормалната работа на системата, т.е. целта на нарушителя е да извърши непозволено и неоторизирано действие. Те се основават на предварително дефинирани профили на нормалната работа на потребителите в системата [4]. Ако текущите събития се отклоняват в значителна степен от тези профили, системата издава съобщение за атака [10]. Основното предимство на системите, базирани на аномалии, е потенциалът им да откриват нови или неизвестни атаки, вариации на съществуващи атаки, както и отклонения от нормалната работа на процесите независимо дали източникът на отклоненията е оторизиран потребител или външно лице. Те могат успешно да открият кражба на потребителски акаунт, тъй като за атакуващия е много трудно да знае със сигурност кои негови действия биха предизвикали алармен сигнал [6]. Недостатъкът на този подход е в сложността на получените системи, както и в обстоятелството, че те са критично зависими от профилите на нормалните потребителски действия, поради което могат да не открият дори добре известни атаки, ако те съответстват на тези профили. Друг недостатък на тези системи е в необходимостта от време за натрупване на достатъчно данни за съставяне на потребителските профили, като през този период системата е незащитена. Съществуват много проблеми при избор на технология за създаване на системи за отчитане на нарушения: ниска производителност, висок процент на фалшиви аларми, висок процент на неправилно класифицирани състояния и др.

Настоящият доклад се фокусира върху хост-базираните системите, основани на аномалии, чиято цел е да създадат описание на нормалната работа на легитимните потребители на системата. След това тези системи следят текущите данни за потребителска активност в системата и търсят отклонение от така дефинираните профили. В случай, че е засечено значително отклонение от тях, системата издава съобщение за нарушение. В предложената методология описанието на профилите на нормалните потребителски действия чрез използване на дърво на решенията. За класифициране на текущите събития в системата като нормални или злонамерени е приложена клъстеризация като данните се разделят на два клъстера – един за нормалните и един за данните, резултат от неоторизирани действия.

2. Методология

Алгоритъмът C4.5, разработен от Рос Куинлан [5], е алгоритъм за генериране на дърво на решенията и често се нарича статистически класификатор, тъй като се използва за класифициране. Той е разширение на преди това предложения от него алгоритъм ID3 (Iterative Dichotomiser 3). В настоящата работа този метод се използва, за да се изгради ефективно дърво на решенията за откриване на проникванията в системата, който след това ще се използва при класифициране на състоянията на системата като нормални или резултат от неоторизирани действия. Предимствата на алгоритъма са, че изгражда бързо дърво на решенията, изгражда късо дърво на решенията, и лесно могат да се определят правила да прогнозиране от тестовите данни. Множеството от тестовите данни е множество от вече класифицирани извадки $S=s_1,s_2,...$. Всяка извадка s_i е p-размерен вектор $\left(x_{1,i},x_{2,i},...,x_{p,i}\right)$, където всяка координата е елемент от описваните данни, както и класа, в който s_i попада. Като възли на дървото за алгоритъма са избрани координатите, които в най-добра степен разделят множеството S на подмножества. Като класификационен критерий се използва получената информация I (information gain). Координатата с най-високата получена информация се избира за възел на дървото за вземане на решение. Тя се пресмята като разлика на ентропията на възела родител със средната ентропия на възлите наследници.

Алгоритъмът за изграждане на дървото на решенията е следният:

- Проверява се за базови класове;
- За всяка координата се пресмята получената информация I за разбиване на подмножества;
- Нека a е координатата с най-голяма I. Приема се като възел за вземане на решения, който ще разбие множеството;
- Рекурсивно в подсписъците, получени от това разделяне се повтаря процедурата, като новополучените възли се добавят в дървото като наследници на предходните.

В [7] процесът на откриване на аномалии се разглежда като двоичен класификационен проблем и за откриване на аномалии е използван метод, основаващ се на 2-means клъстерен алгоритъм. Разстоянията до клъстерния центроид в клъстера се изчисляват с помощта на разстояние на Демерау-Левенщайн [8], [9].

Много изследвания в последните години показват, че хибридните подходи при изграждане на системи за отчитане на нарушния водят до подобряване на точността и надежнастта. Поради тази причина вниманието се фокусира върху хибридни методологии за комбиниране на предимствата на два алгоритъма.

В настоящата работа се комбинират 2-means клъстерен алгоритъм и алгоритъмът C4.5. Първо, прилага се 2-means клъстерен алгоритъм за групиране на тестовите данни [7]. След това във всеки клъстер се изгражда дърво на решениета с помощта на алгоритъма C4.5. Първият алгоритъм разделя данните на клъстери, а вторият прецизира разделянето в клъстерите.

Съществуват много методи за оценяване на най-добрия начин за разбиване на множеството на подмножества. Нека с $p\left(i|t\right)$ се означи вероятността координата да принадлежащи на клас i при даден възел t. Като мярка за разделянето може да се използват следните индекси:

$$Gini(t) = 1 - \sum_{i=0}^{c-1} p(i|t)^2,$$

$$Classification \quad error(t) = 1 - \max_i p(i|t),$$

където *c* е броят на класовете. Индексът Gini е мярка за честотата на попадане на координатата при разделянето на множеството в неправилното подмножество. Той достига своя минимум (нула) когато всеки елемент за даден възел попада в едно подмножество.

Когато се анализира клъстерът, естествено е да се предположи, че клъстерите с по-голям брой вектори представят нормалните състояния, а тези с по-малък брой – аномалиите, както и че векторите в един клъстер са близки по разстояние. Но в случай на мащабна атака, в голям процент от нормалните вектори се получават аномалии. Поради тази причина са налага да се анализират структурите на клъстерите, за да се постигне по-добра класификация. За тази цел се изчислява размерът и разстоянието между клъстерите, както и клъстерната компактност. Компактността се използва за описание на приликите между обектите в един и същи клъстер. Като мярка за клъстерна компактност се използва вътреклъстерно разстояние (*Intra-cluster distance*). То е малко, когато обектите са близко до клъстерния центроид и се увеличава при намаляване на броя на клъстерите. Един от начините за пресмятането му е като диаметър, който е най-голямото разстояние между два вектора в клъстъра $\Delta(K_i) = \max_{x,y \in K_i} \{d(x,y)\}$.

Като мярка за клъстерното разделение се използва междуклъстерното разстояние (*Inter-cluster distance*). Пресмята се по един от следните начини:

- Като единична връзка, която е най-близкото разстояние между две наблюдения, принадлежащи на два различни клъстъра K_i и K_j: $\delta(K_i, K_j) = \min\left\{ \underset{x \in K_i, y \in K_j}{d}(x,y) \right\}$. Големите стойнсти показват по-добра отдаленост между центровете на клъстерите, а малката стойност – че има малък брой големи клъстера;
- Като цялостна връзка, която е най-отдалеченото разстояние между две наблюдения, принадлежащи на два различни клъстъра K_i и K_j: $\delta(K_i, K_j) = \max\left\{ \underset{x \in K_i, y \in K_j}{d}(x,y) \right\}$.

Един от методите на валидност, чрез които се оценява компактността на клъстърите и разстоянията между тях е Индекс на Дън [3] се дефинира като частно на минималното междуклъстърно разстояние и максималното вътреклъстърно разстояние

$$D = \frac{\delta_{\min}}{\Delta_{\max}}$$

Този индекс се ограничава в интервала $[0, \infty)$ и трябва да се максимизира.

Ние се ограничаваме до проучване на случая с два клъстера, единият от които съответства на нормалната дейност, а другият на нарушенията. Логиката на този подход е предположението, че нормална дейност и аномалиите формират различни клъстери. Векторите на атаките често много си приличат, ако не са идентични. Очакването е, че клъстера на атаките в случай на масирана атака е изключително компактен.

3. Оценка на приложената методология

Целта на системите за откриване на нарушения е да се определи дали дадена последователност от наблюдения принадлежи към една от двете групи – множество от нормални дейности на системата или множество от нарушения. За всяко възможно наблюдение тестът, чрез който се реализира тази класификация, може да допуска два типа грешки – грешка от първи род (*false positive – FP*) и грешка от втори род (*false negative – FN*). *FP* се допуска, когато едно събитие се отчита като нарушение, но всъщност това е нормална дейност, докато *FN* е грешката, която се допуска, когато настъпва нарушение, но то не е класифицирано като такова. *TP* (*true positive*) и *TN* (*true negative*) са коректно класифицираните нормални действия и нарушения съответно. При двоичната класификация са възможни четирите резултата. Оценката на ефективността и настройката на системата за откриване на нарушения се нуждае от баланс между тези четири стойности. За да се направи оценка на ефективността на предложената методология, използваща разстоянието на Демерау-Левенщайн, са приложени някои статистически методи. Като показатели за измерване на точността на класификацията са използвани процент на грешки и точност.

Точност (*Accuracy*) е степента на съответствие на брой открити аномалии от метода при дадена непозната последователност с реалния брой аномалии в данните [2].

$$Точност = \frac{TP + TN}{Брой \quad на \quad входящи \quad вектори}$$

Колкото по-висока е стойността на точността, толкова тестът идентифицира нарушения и нормална активност с по-висока прецизност.

Процент на грешките (*Error rate*) се изчислява със следната формула:

$$Процент \quad на \quad грешките = \frac{FP + FN}{Брой \quad на \quad входящи \quad вектори}.$$

Класификационните алгоритми се стремят да постигнат висока точност или еквивалантно нисък процент на грешките.

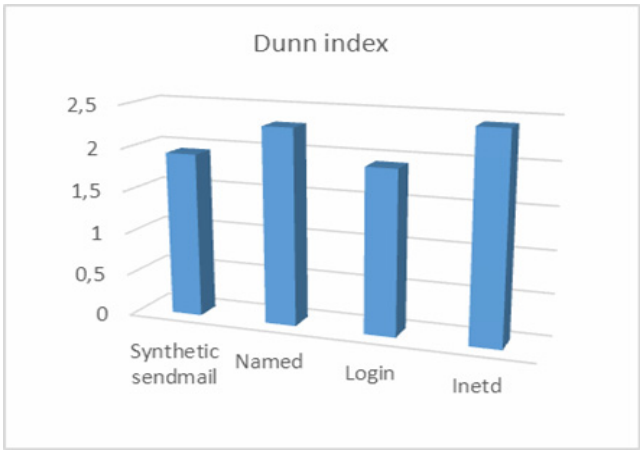
4. Симулационни изследвания

Предложената методология е тествана чрез симулационни експерименти, за които са използвани данни за процеси (synthetic ftp, xlock, login, named, synthetic lpr), изпълнявани с администраторски права в Unix система [11]. Разгледаните набори от данни включват ID на изпълнявания процес и номера на системното извикване:

PID	1393	1393	...	1423
Системно извикване	112	19	...	105

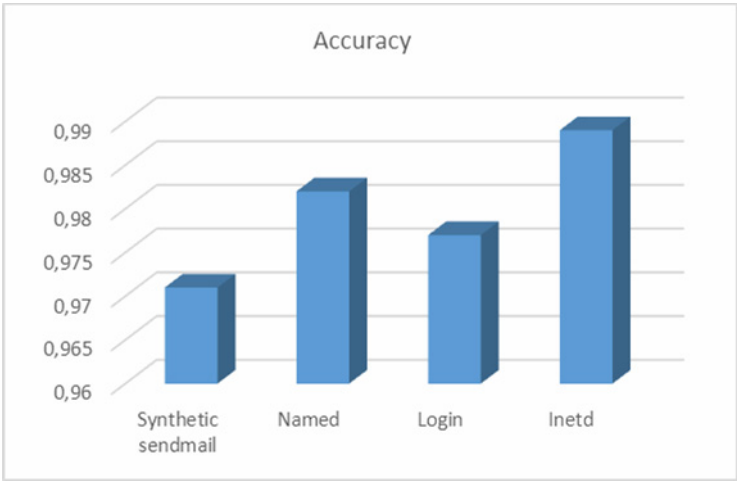
Таблица 1. Примерни тестови данни

При направените експерименти стойностите на мярката за разделяне Gini индекс са в интервала (0,387; 0,462). На графиката на фигура 1 са представени стойностите на индекса за валидност Дън. Основната цел на мярката е да се определи дали предложеният метод постига максимизиране на вътреклъстърното разстояние и свеждане до минимум на междуклъстърно разстояние. От фигура 1 е ясно, че при направения експеримент са постигнати добри резултати от гледна точка на индекса Дън.



Фигура 1. Стойности на Индекса на Дън от проведения експеримент

Експериментите показват, че процентът на точност на този алгоритъм за откриване на ненормални състояния е над 0, 97 както се вижда от графиката на фигура 2:



Фигура 2. Стойности на точността от проведения експеримент

5. Заключение

В настоящия доклад е представена методология, реализираща описание на нормалните потребителски действия в даден хост, включващ действията, извършвани от привилегировани процеси чрез дърво на решенията. Сканирането на данните, описващи текущата работа в системата и класифицирането им като нормални или резултат от злонамерени действия се извършва чрез клъстериране, като се използват разстояния между последователности от елементи.

Предмет на бъдещи изследвания може да бъде приложението на предложения метод върху мрежови данни, както и сравнението му с други методи, реализиращи система за откриване на нарушения, основаваща се на поведенчески анализ.

Литература

1. Abraham, A., Thomas, J., Distributed Intrusion Detection Systems: A Computational Intelligence Approach, Applications of Information Systems to Homeland Security and Defense, Idea Group Inc. Publishers, USA, Chapter 5, 2005, pp. 105-135.
2. Baldi P., Brunak S., Chauvin Y., Andersen C.A., Nielsen H., Assessing the accuracy of prediction algorithms for classification: An overview, Bioinformatics, Vol. 16, pp. 412–424, 2000.
3. Dunn, 1974. Dunn, J. (1974) Well separated clusters and optimal fuzzy partitions, Journal of Cybernetics , 4, 95-104.
4. Qiao Y., X.W. Xin, Y. Bin, S. Ge, „Anomaly intrusion detection method based on HMM”, IEEE Electronic Letters Online No: 20020467, 2002.
5. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.
6. Michael C. C., A. Ghosh, „Simple, state-based approaches to program-based anomaly detection”, ACM Transactions on Information and System Security, Vol. 5, No. 3, August 2002, pp. 203-237.
7. Nikolova E., V.Jecheva, An Adaptive Approach of Clustering Application in the Intrusion Detection Systems, OPEN JOURNAL OF INFORMATION SECURITY AND APPLICATIONS, Volume 1, Number 3, December 2014.
8. Damerau F. J., „A technique for computer detection and correction of spelling errors,” Communica-tions of the ACM, vol. 7, no. 3, pp. 171–176, 1964.
9. Levenshtein V. I., „Binary codes capable of correcting deletions, insertions and reversals,” in SovietPhysics Doklady, vol. 10, p. 707, 1966.
10. Tran T.P., T. Jan, A.J. Simmonds, A Multi-Expert Classification Framework for Network Misuse Detection, From Proceeding (544) Artificial Intelligence and Soft Computing, ISBN 0-88986-610-4, 2006.
11. University of New Mexico’s Computer Immune Systems Project, http://www.cs.unm.edu/~immsec/systemcalls.htm.