N-Gram 증강 나이브 베이스를 이용한 정확한 침입 탐지

강대기

동서대학교

Accurate Intrusion Detection using n-Gram Augmented Naive Bayes

Dae-Ki Kang

Dongseo University

E-mail : dkkang@dongseo.ac.kr

요 약

기계 학습을 응용한 많은 침입 탐지 시스템들은 n-그램 접근 방법을 주로 쓰고 있다. 그러나, n-그램 접근 방법은 주어진 시퀀스에서 획득한 n-그램들이 서로 겹치는 문제들을 가지고 있다. 본 연구에서는 이러한 문제들을 해결하기 위해, n-그램 증강 나이브 베이스 (n-gram augmented naive Bayes) 알고리즘을 침입 시퀀스의 분류에 적용하였다. 제안된 시스템의 성능을 평가하기 위해 n-그램 특징들을 사용하는 일반 나이브 베이스 (naive Bayes) 알고리즘과 서포트 벡터 머신 (support vector machines) 알고리즘과 본 연구에서 제안한 n-그램 증강 나이브 베이스 알고리즘을 비교하였다. 뉴 멕시코 대학의 벤치마크 데이터에 적용해 본 결과에 따르면, n-그램 증강 방법이, n-그램이 나이브 베이스에 직접 적용되는 경우(예: n-그램 특징을 사용하는 일반 나이브 베이스), 생기는 독립성 가정에 대한 위배 문제도 해결하면서, 동시에 n-그램 특징을 사용하는 일반 나이브 베이스보다 더 정확하며, n-그램 특징을 사용하는 SVM과 필적할만한 수준의 침입 탐지기를 생성해 내었다.

ABSTRACT

In many intrusion detection applications, n-gram approach has been widely applied. However, n-gram approach has shown a few problems including double counting of features. To address those problems, we applied n-gram augmented Naive Bayes directly to classify intrusive sequences and compared performance with those of Naive Bayes and Support Vector Machines (SVM) with n-gram features by the experiments on host-based intrusion detection benchmark data sets. Experimental results on the University of New Mexico (UNM) benchmark data sets show that the n-gram augmented method, which solves the problem of independence violation that happens when n-gram features are directly applied to Naive Bayes (i.e. Naive Bayes with n-gram features), yields intrusion detectors with higher accuracy than those from Naive Bayes with n-gram features and shows comparable accuracy to those from SVM with n-gram features.

키워드

N-그램 나이브 베이스 알고리즘, 침입 탐지

Ⅰ. 서 론

데이터 마이닝 알고리즘은 호스트 기반 침입 탐지 작업에서 프로그램의 트레이스(program trace)를 분류하는 데 널리 사용되어 왔다. 구체적으로 침입 탐지 작업에서 데이터의 전처리로서, 시스템 콜 트레이스에서의 특징 추출을 위해 n-그램(프로그램 트레이스 내에서 n 개의 연속된 시스템 콜) 방법[1]이 널리 사용되어 왔다[2-4]. 그러나, 이러한 n-그램 접근 방법은 침입 탐지에 적용되기에는 세 가지 심각한 문제점을 안고 있다.

1. 운영 체제에서 시스템 콜의 개수는 약 200 여 개이므로, n-그램 방식의 특징들의 개수는 n이 증가하면 빠르게 증가한다. 예를 들면, 뉴 멕시코 대학의 벤치마크 데이터로 사용된 SunOS의 시스템 콜의 개수는 183 개인데, 만일 20-그램이

사용되었다면 그 개수는 1,774,278,518,944,245,232,888,176,323,498,992,582,562,189,601이나 되므로, 실제 응용에는 실용적이지 못하다.

2. n-그램 특징들은 고정된 크기의 윈도우를 사용하여 원래의 프로그램 트레이스로부터 생성된다. 따라서 프로그램 트레이스 내의 하나의 특정 시스템 콜이 적어도 n 개의 특징들 내부에 포함될 수 있다[5].
3. 만일 생성된 침입 탐지 시스템이, 예를 들면 나이브 베이스 알고리즘과 같이, 특징들 간의 통계적인 독립성에 대한 가정에 의지한다면, 2에서 언급한 n-그램 특징 생성 방법은 근본적으로 이러한 가정을 위배한다.

서포트 벡터 머신 (SVM)[6,7]과 같이 비선형적인 복잡도를 가지는 데이터 마이닝 알고리즘은 첫 번째 문제에 취약하다. 두 번째와 세 번째 문제를 해결하기 위해 텍스트 분류 및 바이오인포매틱스 분야에서는 n-그램 증강 나이브 베이스 기법이 사용되어왔으나[5,8,9], 침입 탐지 분야에서는 이 기법이 연구된 바 없다.

이러한 배경으로, 우리는 n-그램 증강 나이브 베이스를 호스트 기반 침입 탐지 작업에 적용하였고, 그 성능을 n-그램 특징을 사용하는 나이브 베이스와 n-그램 특징을 사용하는 SVM과 비교하였다.

호스트 기반 침입 탐지 벤치마크 데이터에 대해 행한 실험 결과에 따르면, 본 연구에서 응용한 n-그램 증강 나이브 베이스가 n-그램 특징을 사용하는 나이브 베이스보다 더 좋은 결과를 보였으며, n-그램 특징을 사용하는 SVM와 비슷한 정확도를 보였다.

II. 본 론

우리는 우선, n-그램 특징을 사용하는 나이브 베이스 (NB n-gram)과 n-그램 증강 나이브 베이스, 그리고 n-그램 특징을 사용하는 SVM (SVM n-gram)에 대해 설명하고자 한다. 각 방법들을 설명하기 전에, 호스트 기반 침입 탐지 문제를 형식적으로 정의해보고자 한다.

$\Sigma = \{s_1, s_2, s_3, s_4, \ldots, s_m\}$을 시스템 콜들의 집합이라고 할 때, 데이터 집합 D는 레이블이 붙은 시퀀스(즉 트레이스)의 집합들로 $D = \{\langle Z_i, c_i \rangle | Z_i \in \Sigma^*, c \in \{0,1\}\}$로 정의될 수 있다. 여기서, $Z_i = z_1, z_2, z_3, \ldots, z_l$는 입력 시퀀스이고, c_i는 이 입력 시퀀스에 상응하는 클래스 레이블로, 0은 침입이 아님을 뜻하고, 1은 침입을 뜻한다. 이렇게 데이터 집합 D가 주어지면, 침입 탐지 학습 알고리즘의 목표는 정확도와 같은 주어진 평가 기준을 최대화하는 침입 탐지기 $h : \Sigma \rightarrow 0,1$를 발견하는 것이다.

만일 나이브 베이스와 같은 확률적인 모델을 침입 탐지기 h에 사용한다면, 결과적인 확률 모델 P_h는 주어진 시퀀스 Z 에 대해 다음과 같이 확률 $P_h(Z = z_1, z_2, z_3, \ldots, z_l)$를 설정한다.

- 각 클래스 c_i에 대해, c_i와 연관된 시퀀스들을 샘플링하여 확률 $P_h(c_i)$를 추정함
- 새로운 시퀀스 Z에 대하여 클래스 c를 다음에 근거하여 설정함

$$c_h = \arg max_{c \in \{0,1\}} P_h(Z = z_1, z_2, \ldots, z_l | c) P_h(c)$$

2-1. 나이브 베이스 분류기

호스트 기반 침입 탐지기로서의 나이브 베이스 분류기의 중요한 가정 중 하나는, 주어진 클래스에 대해 시퀀스의 각 시스템 콜이 서로 독립적이라는 것이다. 그러므로, 나이브 베이스의 경우 새로운 시퀀스에 대한 분류는 다음과 같이 형식화될 수 있다.

$$c_{NB} = \arg max_{c \in \{0,1\}} P_h(c) \cdot \prod_i P_h(z_i | c)$$

2-2. n-그램 특징을 사용하는 나이브 베이스

각 시퀀스들은 길이가 고정되어 있지 않으므로, 고정되고 유한한 크기의 입력을 받는 컴퓨터 알고리즘에 적용하는 데에는 어려움이 있다. 따라서, 주어진 시퀀스는 유한한 크기의 n 차원 특징 벡터(n-그램)로 변환된다.

호스트 기반 침입 탐지 작업에서는 프로그램의 행동을 모니터하게 되고, 이를 위해서 프로그램의 트레이스를 시퀀스로 간주한다. 따라서, n-그램 특징들을 생성하기 위해서, 길이가 n 인 슬라이딩 윈도우를 트레이스의 시작부터 끝까지 시스템 콜 하나씩 옮겨가면서, 주어진 트레이스로부터 n-그램 특징들을 생성해 낸다.

특징 추출이 끝나면, 생성된 n-그램들에 대한 확률적 모델은 다음과 같이 형식화될 수 있다.

$$c_{NB\,n-gram} = \arg max_{c \in \{0,1\}} P_h(c) \cdot \prod_i^{l-n+1} P_h(z_i, \ldots, z_{i+n-1} | c)$$

이러한 n-그램 특징을 사용하는 나이브 베이스 모델은 한 가지 심각한 문제점을 가지고 있다. 그것은 n-그램 특징이 슬라이딩 윈도우를 통해 생성되는 동안, 트레이스 안의 특정 시스템 콜 하나가 많으면 n 번 이상 슬라이딩 윈도우 안에 포함된다는 것이다. 결국 서로 이웃한 n-그램 특징은 구조적으로 서로 독립적이지 않으므로, 나이브 베이스 학습 알고리즘의 독립성 가정에 위배되게 된다.

2-3. n-그램 증강 나이브 베이스

앞서 언급한 문제를 해결하기 위해, Peng과 Schuurmans[8]은 n-그램 증강 나이브 베이스를

도입하여 텍스트 분류 문제에 적용하였다. 그들은 시퀀스로부터 만들어진 n-gram 특징 내부에 있는 요소들 간의 의존성을 명시적으로 모델링하는 접근 방법을 택했다. 그림 1은 이러한 접근 방법으로 시퀀스 내의 여섯 개의 서로 이웃한 요소들의 의존 관계를 표현한 것이다.

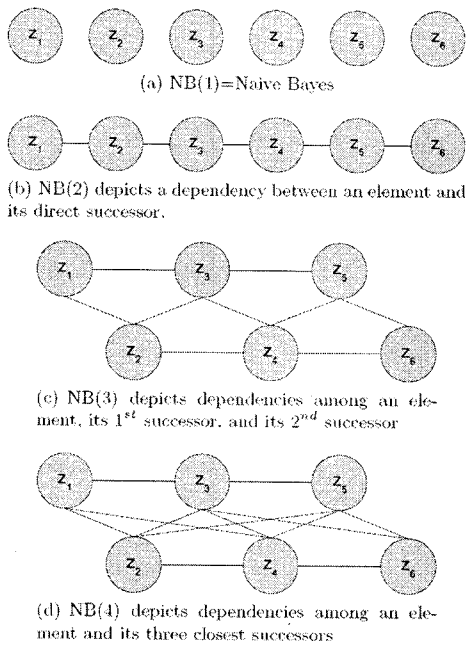

그림 1. 시퀀스 안의 연속된 여섯 개의 요소들 간의 의존도를 표현한 모델

정션 트리 정리[10]에 의해, n-그램 증강 나이브 베이스의 확률적 모델은 다음과 같이 형식화될 수 있다.

$$c_{NB(n)} = \arg max_{c \in \{0,1\}} \frac{\prod_{1}^{l-n+1} P_h(z_i,\ldots,z_{i+n-1}|c)}{\prod_{2}^{l-n+1} P_h(z_i,\ldots,z_{i+n-2}|c)} P_h(c)$$

그림 1과 위의 식에서 알 수 있듯이, n-그램 증강 나이브 베이스의 확률적 그래피컬 모델은, 근본적으로 마르코프 네트워크으로 그 확률 분포는 최대 클리크(clique)들의 마지널(marginal)들의 곱을 세퍼레이터(separator)들의 마지널(marginal)들의 곱으로 나누어서 구해진다.

2-4. n-그램 특징을 사용하는 SVM

n-그램 증강 나이브 베이스의 성능을 다른 데이터 마이닝 알고리즘과 비교해 보기 위해, 우리는 n-그램 특징들을 사용하는 SVM을 고려해보았다. 즉, 우리는 원래의 프로그램 트레이스에서 n-그램 특징들이 구하고, 구해진 특징들은 선형 커널을 사용하는 SVM 알고리즘의 입력으로 사용하였다.

우리의 관심사는 n-그램 증강 나이브 베이스와 n-그램 특징을 사용하는 나이브 베이스를 n-그램 특징을 사용하는 SVM과 비교하는 것인데, 그 이유는 n-그램 특징을 사용하는 나이브 베이스와 달리, SVM은 독립성에 대한 가정에 의존하지 않기 때문이다.

그러나, SVM 알고리즘은 비선형 복잡도를 가지므로, n이 증가함에 따라 필요한 저장 공간은 나이브 베이스보다 더 빠르게 증가하는 문제가 있다. 즉, SVM 의 커널 매트릭스를 준비하는 데만, 적어도 $O(n^2)$ 의 시간 및 공간 복잡도를 가진다.

실제 실험에서는, 이러한 컴퓨팅 및 메모리 문제로, 우리는 SVM을 n 이 1 또는 2 인 경우에 대해서만 수행할 수 있었다.

III. 실험 및 결과

n-그램 증강 나이브 베이스의 성능을 평가하기 위해, 우리는 그 성능을 n-그램 특징을 사용하는 나이브 베이스와 n-그램 특징을 사용하는 SVM과 비교하였다. 실험을 위한 데이터로 공개적으로 사용가능한 뉴 멕시코 대학(University of New Mexico)의 "*UNM live lpr*"시스템 콜 트레이스들을 사용하였다.

표 1은 이러한 세 개의 알고리즘의 결과로 나온 정확도(accuracy)와 거짓 양성(false positive) 값을 나타낸 것이다. 결론부터 말하면, n-그램 증강 나이브 베이스는 n이 6에서 8일 때 가장 좋은 성능을 보였다. 보여준 성능은 SVM의 경우와 필적할만 했다.

표 1. 해양정보통신의 영역

n	NB(n)		NB n-그램		SVM n-그램	
	A	FP	A	FP	A	FP
1	84.09	28.84	84.09	28.84	100.0	0.00
2	99.78	0.41	98.30	3.09	99.96	0.00
3	99.96	0.00	99.01	1.79	N/A	N/A
4	99.96	0.00	99.60	0.73	N/A	N/A
5	99.96	0.00	99.82	0.32	N/A	N/A
6-8	100.0	0.00	99.87	0.24	N/A	N/A
9-10	99.96	0.08	99.82	0.32	N/A	N/A
11-20	99.96	0.08	99.78	0.41	N/A	N/A

전체적으로 n-그램 증강 나이브 베이스는 n-그램 특징을 사용하는 나이브 베이스보다 더 나은 성능을 보였다. 표 1의 "UNM live lpr" 데이터

를 보면, n-그램 증강 나이브 베이스와 n-그램 특징을 사용하는 나이브 베이스는 둘 다 n 이 6에서 8일때, 최적의 성능을 보였다. n-그램 증강 나이브 베이스가 보인 최고의 정확도와 거짓 양성율은 100.00과 0.00이고, n-그램 특징을 사용하는 나이브 베이스의 최고의 정확도와 거짓 양성율은 99.87과 0.24이었다.

IV. 결론

본 연구에서는 n-그램 증강 나이브 베이스 (n-gram augmented naive Bayes) 알고리즘을 침입 시퀀스의 분류에 적용하였다. 제안된 시스템의 성능을 평가하기 위해 n-그램 특징들을 사용하는 일반 나이브 베이스 (naive Bayes) 알고리즘과 서포트 벡터 머신 (support vector machines) 알고리즘과 본 연구에서 제안한 n-그램 증강 나이브 베이스 알고리즘을 비교하였다. 뉴 멕시코 대학의 벤치마크 데이터에 적용해 본 결과에 따르면, n-그램 증강 방법이, n-그램이 나이브 베이스에 직접 적용되는 경우(예: n-그램 특징을 사용하는 일반 나이브 베이스), 생기는 독립성 가정에 대한 위배 문제도 해결하면서, 동시에 n-그램 특징을 사용하는 일반 나이브 베이스보다 더 정확하며, n-그램 특징을 사용하는 SVM과 필적할만한 수준의 침입 탐지기를 생성해 냄을 알 수 있었다.

차후 연구 방향으로, 우리는 DARPA 1998/1999 데이터[11]와 같은 더 많은 데이터 집합들에 대해 실험을 확장할 계획이다. 가능한 다른 연구 방향으로는 n-그램 표현을 시스템 콜의 매개 변수들에 대해 확장하는 것[12]이다.

참고문헌

[1] E. Charniak, Statistical Language Learning, MIT Press, Cambridge, MA, USA, 1994.

[2] W. Lee, S. J. Stolfo, K. W. Mok, A data mining framework for building intrusion detection models, in: IEEE Symposium on Security and Privacy, 1999, pp. 120-277 132.

[3] A. Murali, M. Rao, A survey on intrusion detection approaches, in: First International Conference on Information and Communication Technologies (ICICT 2005), 2005, pp. 233-240.

[4] K. Rieck, P. Laskov, Detecting unknown network attacks using language models., in: Proceedings of Third International Conference on Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA 2006), Berlin, Germany, 2006, pp. 74-90.

[5] C. Andorf, A. Silvescu, D. Dobbs, V. Honavar, Learning classifiers for assigning protein sequences to gene ontology functional families, in: Proceedings of the Fifth International Conference on Knowledge Based Computer Systems (KBCS 2004), 2004, pp. 256-265.

[6] B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal 291 margin classifiers, in: COLT '92: Proceedings of the fifth annual workshop on 292 Computational learning theory, ACM Press, New York, NY, USA, 1992, pp. 293 144-152.

[7] V. N. Vapnik, The nature of statistical learning theory, Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[8] F. Peng, D. Schuurmans, Combining naive Bayes and n-gram language models for text classification., in: F. Sebastiani (Ed.), Advances in Information Retrieval, 25th European Conference on IR Research (ECIR 2003), Vol. 2633 of Lecture Notes in Computer Science, Springer, 2003, pp. 335-350.

[9] A. Silvescu, C. Andorf, D. Dobbs, V. Honavar, Inter-element dependency models for sequence classification, Tech. rep., Iowa State University (June 2004).

[10] R. G. Cowell, S. L. Lauritzen, A. P. David, D. J. Spiegelhalter, D. J. Spiegelhater, Probabilistic Networks and Expert Systems, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1999.

[11] R. Lippmann, R. K. Cunningham, D. J. Fried, I. Graf, K. R. Kendall, S. E. Webster, M. A. Zissman, Results of the DARPA 1998 offline intrusion detection evaluation, in: Recent Advances in Intrusion Detection, 1999.

[12] D. Mutz, F. Valeur, G. Vigna, C. Kruegel, Anomalous system call detection, ACM Trans. Inf. Syst. Secur. 9 (1) (2006) 61-93.