네트워크 침입 탐지를 위한 최적 특징 선택 알고리즘

정승현* · 문준걸* · 강승호*

*동신대학교 정보보안학과

An optimal feature selection algorithm for the network intrusion detection system

Seung-Hyun Jung* · Jun-Geol Moon* · Seung-Ho Kang*

*Dept. of Information Security, Dongshin University

E-mail : {tmdgus7923, jkloy}@naver.com, drminor@dsu.ac.kr

요 약

기계학습을 이용한 네트워크 침입탐지시스템은 선택된 특징 조합에 따라 정확성 및 효율성 측면에서 크게 영향을 받는다. 하지만 일반적으로 사용되는 침입탐지용 특징들로부터 최적의 조합을 찾아내는 일은 많은 계산량을 요구한다. 예를 들어 n개로 구성된 특징들로부터 가능한 특징조합은 2^n-1 개이다. 본 논문에서는 이러한 문제를 해결하기 위한 최적 특징 선택 알고리즘을 제시한다. 제안한 알고리즘은 최적화 문제 해결을 위한 대표적인 메타 휴리스틱 알고리즘인 지역탐색 알고리즘에 기반 한다. 또한 특징 조합을 평가를 위해 선택된 특징 요소와 k-means 군집화 알고리즘을 이용해 구해진 군집화의 정확성을 비용함수로 사용한다. 제안한 특징 선택 알고리즘의 평가를 위해 NSL-KDD 데이터와 인공 신경망을 사용해 특징 모두를 사용한 경우와 비교한다.

ABSTRACT

Network intrusion detection system based on machine learning methods is quite dependent on the selected features in terms of accuracy and efficiency. Nevertheless, choosing the optimal combination of features from generally used features to detect network intrusion requires extensive computing resources. For instance, the number of possible feature combinations from given *n* features is 2^n-1. In this paper, to tackle this problem we propose a optimal feature selection algorithm. Proposed algorithm is based on the local search algorithm, one of representative meta-heuristic algorithm for solving optimization problem. In addition, the accuracy of clusters which obtained using selected feature components and k-means clustering algorithm is adopted to evaluate a feature assembly. In order to estimate the performance of our proposed algorithm, comparing with a method where all features are used on NSL-KDD data set and multi-layer perceptron.

키워드

network intrusion detection system, machine learning, feature selection, local search algorithm

Ⅰ. 서론

다양한 컴퓨팅 기기의 등장과 유무선 통신망의 확대는 대규모의 정보 공유를 가능하게 하였고 다양한 분야에서 효율성과 생산성을 높이는데 기여하였다. 하지만 이에 수반해 악의적인 사용자들에 의한 정보 침해 사고 또한 빈발하고 있으며 통신망의 발전에 많은 걸림돌이 되고 있다. 이러한 문제의식 하에 네트워크에서 이루어지는 악의적 사용자에 의한 비정상적 사용을 자

동으로 탐지해내고 방지하려는 기계학습 기반 침입탐지 시스템에 대해 연구자들의 관심이 높으며 이와 관련한 많은 연구가 진행되고 있다 [1, 2].

기계학습에 기반한 침입탐지 시스템의 개발과 관련해 가장 중요한 요소 중 하나는 침입과 정상적 사용을 표현하고 구분이 가능하도록 하는 특징이다. 다양한 특징들이 제시되고 있으나 공개된 데이터의 부족으로 제안된 특징들의 객관적이고 공정한 평가를 하기에 많은 제약이 되고 있다. 이러한 단점들 극복하고 제안된 침입 탐지 방법들의 객관성과 공정성을 높이기 위해 KDD'99 데이터 집합이 제공되었다 [3]. 이후 많은 연구자들이 KDD'99 데이터를 대상으로 다양한 탐지 방법들을 제안하였다. 하지만 KDD'99 데이터는 총 41개의 특징 요소를 사용하고 있어서 실시간 침입 탐지에 사용하기에는 특징 벡터의 크기가 너무 크다는 비판이 제기되었다. 이러한 문제를 해결하기 위해 중요한 특징 요소만을 선택하려는 연구들이 최근 관심을 모으고 있다 [4]. 그러나 제안된 방법들 대부분은 개별적 특징 요소의 정보 획득(information gain) [5] 이나 의존성 비율(dependency ratio) [6], 상관계수 [7] 등과 같은 상관성을 분석하거나 특성에 순위를 매겨 제거하는 방식 [8] 등에 의존하고 있다. 하지만 특징 요소들의 조합적 특성은 개별 특징 요소들의 단순 합과는 다른 창발적(emergent) 효과를 시스템에 가져올 수 있다는 점에서 이러한 접근 방법에 많은 문제가 있음을 알 수 있다. 특징선택을 위한 기존 연구들이 이러한 접근 방법의 한계를 인식하면서도 사용하는 가장 큰 이유는 대상이 되는 특징 조합의 수가 너무 많아 실험적으로 해결 불가능하기 때문이다. 즉 n개의 특징 요소로부터 선택 가능한 특징 조합은 $2^n - 1$ 가지 경우나 되기 때문이다.

본 논문은 이러한 조합 문제를 해결하기 위한 방법을 제시하고자 한다. 이를 위해 대상이 되는 공격의 범주는 서비스 거부 공격(DoS: Denial of Service)에 한정 하고 그 종류에 상관없이 DoS 공격 여부만을 판별하는 침입탐지 시스템용 최적 특징 추출 알고리즘을 제안한다. 지금까지 특징 선택 방법들과는 달리 특징 선택 문제를 조합 최적화의 관점에서 정의하고 조합 최적화 문제를 해결하기 위해 메타 휴리스틱 알고리즘 중 하나인 지역탐색 알고리즘을 제안한다. 그리고 탐색 알고리즘의 성능을 좌우할 비용 함수를 위해 선택된 특징 조합만을 사용해 군집화하고 이의 정확성을 비용함수로 사용하는 방법을 설계하였다. 제안한 특징 선택 알고리즘의 성능을 평가하기 위해 인공 신경망을 설계하고 41개 특징 요소를 모두 사용한 경우와 선택된 특징만을 사용한 경우의 정확성을 [9]에 의해 제공된 NSL_KDD 데이터를 사용해 비교하였다.

Ⅱ. 데이터 집합

제안한 특징 추출알고리즘의 성능 측정을 위해 NSL_KDD 데이터 집합 [10]을 사용하였다. NSL_KDD 데이터 집합은 KDD'99 데이터 집합이 가지고 있는 문제점들을 보완한 것으로 다음과 같은 차이가 있다.

- KDD 훈련 데이터나 테스트 데이터가 가지고 있는 중복성을 제거하여 시스템이 빈도가 높은 데이터에 편향(bias)되는 현상을 방지하였다.
- 각 부류의 데이터 집합을 난이도에 따라 크기를 정함으로써 성능 평가에 객관성을 높일 수 있다.
- 훈련 및 테스트 데이터의 수를 적정하게 제한함으로써 전체 집합을 대상으로 실험이 가능하도록 하여 부분 집합 선택 시에 발생하는 자의성을 제거하여 실험 결과끼리의 공정한 비교를 가능하게 하였다.

NSL_KDD 데이터 집합은 KDD'99 데이터 집합과 마찬가지로 공격 유형이 네 가지 카테고리(DoS, R2L, U2R, probing)로 분류되어 있으며 각 데이터 사례는 41개의 특징으로 구성되어 있다. 본 논문은 서비스 거부 공격(DoS)만을 대상으로 필요최적의 특징 집합을 추출하는 것을 목적으로 하고 있기 때문에 전체 데이터에서 정상인 경우와 DoS 공격에 해당하는 데이터만을 따로 추출하여 사용하였다. 이렇게 추출된 데이터의 크기는, 우선 훈련 데이터의 경우 113271개 이고 평가 데이터는 15452개로 구성되어있다. 한편 NSL_KDD 데이터 DoS 공격의 유형이 총 5 가지이며 정상인 경우를 포함함 데이터 구성은 아래 표1과 같다.

표 1. 정상 데이터와 서비스 거부 공격 데이터의 구성

	normal	neptune	teardrop	smurf	pod	back	land
훈련	67344	41214	892	2646	201	956	18
평가	9711	4657	12	665	41	359	7

Ⅲ. 특징 추출 알고리즘

3.1 데이터 전처리

NSL_KDD 데이터가 보유한 특징에는 protocol_type, servie, flag와 같이 숫자가 아닌 특징이 존재하는 한편 단위가 10억이 넘는 src_bytes, dst_bytes와 같은 특징이 존재한다. 이러한 특징들은 숫자 데이터로 변환이 필요하고 특정 특징에 의해 분류기가 왜곡되지 않도록 하기 위해 정규화를 하였다. [2]의 논문을 참조하여 다음과 같은 방법으로 정규화 하였다.

- 심볼릭 특징 - 각 특징이 가지고 있는 종류에 0부터 양의 정수를 부여하고 이를 [0, 1]로 선형 정규화

• src_bytes, dst_bytes와 같이 아주 큰 값을 갖는 특징 : 10을 베이스로하는 로그값으로 정규화
• 이진 값을 갖는 특징 : 0과 1을 그대로 사용
• 나머지 : [0, 1] 사이 값으로 선형 정규화

참고로 정규화 이전의 데이터가 가지고 있는 각 특징의 최소값과 최대값은 다음과 같다.

최소값

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0, 0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]

최대값

[54451, 2, 64, 10, 89581520, 7028652, 1, 3, 3, 101, 4, 1, 7479, 1, 2, 7468, 100, 2, 9, 0, 1, 1, 511, 511, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 255, 255, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]

3.2 특징 선택 알고리즘

본 논문은 기존에 제안된 특징 선택 알고리즘과 달리 특징 선택을 최적화 문제로 정의하고 이를 해결하기 위한 방법을 제시한다. 41개의 요소로 구성된 특징으로부터 가능한 특징 조합은 2^{41}-1개나 된다. 서론에서도 언급했듯이 이러한 이유로 기존에 제시된 대부분의 알고리즘들이 개별 특징 요소들을 대상으로 데이터와의 연관성 고려에 국한 되었다.

이러한 문제를 극복하고자 본 논문은 조합 최적화 문제에 자주 사용되는 지역탐색 알고리즘을 특징 선택 문제에 적용한다. 우선 지역탐색 알고리즘에 사용되는 해는 길이가 41인 2진 벡터로 표현된다 (표 4). 선택된 특징 요소는 1로 그렇지 않는 특징 요소는 0으로 표현된다. 예를 들어 41개의 모든 특징을 사용하는 경우엔 1이 41개인 벡터를 해로 갖는다. 한편, 지역탐색 알고리즘과 같은 최적화 휴리스틱 알고리즘의 성능을 크게 좌우하는 것은 개별 해에 대한 비용함수이다. 즉, 비용함수를 어떻게 정의하느냐에 알고리즘의 성능이 크게 영향을 받는다. 본 논문은 개별 해에 대한 비용함수는 주어진 해를 이용한 군집화의 정확성을 이용한다. 개별 해가 제시하는 특징 조합만을 가지고 학습 데이터를 *k*-means 군집화 알고리즘을 사용해 군집화한 후 학습 데이터를 얼마나 정확하게 분류했는지를 개별해의 비용으로 사용한다. 그리고 탐색의 대상이 되는 특정 해에 대한 이웃 해는 그와 1 비트 다른 해로 정의하였다. 따라서 임의의 해에 대한 이웃 해는 총 41개가 가능하다. 탐색은 이러한 41개의 해 중 가장 정확한 군집을 생성하는 해를 다음 탐색의 대상으로 삼고 동일한 과정을 반복한다. 하지만 어떤 이웃 해도 현재 해의 정확성 보다 높은 정확성을 갖지 못하면 해당 해를 알고리즘의 최적해로 삼고 더 이상의 탐색은 중단한다.

아래 표 2에 특징 선택 알고리즘에 대한 의사코드를 제시한다.

표 2. 특징 추출 알고리즘 의사코드

Algorithm: Feature selection algorithm Input: training data set Output: combination of features
1. Generate an initial solution, S_{init}
2. Calculate the cost of initial solution, cf(S_{init})
3. current best = S_{init}
4. *while*(1)
5. Find neighbor solutions which have one bit different
6. Calculate the cost of each neighbor solution
7. Select the best one from neighbor solutions
8. *if* cf(current best) > cf(best neighbor)
9. *break*
10. *else*
11. current best = best neighbor

IV. 실험 및 분석

제안한 특징 선택 알고리즘의 성능을 평가하고자 선택된 특징을 대상으로 인공 신경망을 사용하여 정확성을 측정하였다. 우선 특징 선택 알고리즘을 사용해 20 개의 해를 구하고 각 해가 제시한 특징 요소만을 사용해 학습 데이터 집합과 평가 데이터 집합을 생성하였다.

인공 신경망의 구조는 3 층 구조로 되어 있다. 입력 층의 노드 개수는 사용하는 특징 요소의 개수와 같고 히든 층의 노드 수는 여러 번의 실험 후 가장 성능이 좋은 노드 수를 사용하였으며 출력 층은 하나의 노드를 사용해 정상인 경우 0을 DoS 공격인 경우 1을 출력하도록 하였다. 학습 알고리즘은 오류 역전파 알고리즘을 사용하였다.

선택된 20개의 특징 조합과 각 특징 조합이 보여준 정확성은 다음 표와 같다.

표 3. 선택된 특징 조합의 정확성

선택된 특징 조합	정확성(%)	
	학습	테스트
00000010110010001000011010111001101100111	99.22	95.59
00010000110111101111000001110011011000111	97.83	93.23
00011000100010010011101011111010000111110	99.14	95.84
00100010011101011111111001011101101011111	98.80	96.51
00100010101100101001111011111001110101001	98.90	96.08
00101010001000111101011111110110000000111	99.94	98.82
00111001001011001111111011111111100000111	99.97	99.33
00111001101001011010111011101101000011111	99.98	99.22
00111011110010001101101111111010001101110	99.98	99.37
01001010101100010011011011111101000011111	99.89	98.67
01100011111011100000101010111000100100111	99.08	96.95
10010010101111001101011001011010011101111	98.54	93.89
10100000001111101101001001100011101010111	99.04	95.32
10101011010011110110001111111000001101110	99.98	99.15
10111001110000011100111011111100000001111	99.97	99.09
11000000110001000111001100011010000101010	99.69	97.97
11000011010000001011110001111000100000101	99.39	98.63
11010001110111010001100010101010010101000	98.67	96.82
11100010000011110010001001100010110101011	99.49	97.39
11111010100010010111111111110111000001101	99.98	98.29

표 3으로부터 제안한 특징 선택 알고리즘을 이용해 얻은 특징 조합을 사용한 경우 학습 데

이터를 대상으로 평균 99.37%의 정확성을, 그리고 테스트 데이터를 대상으로는 97.31%의 정확성을 보여주었다. 한편 41개 모든 특징 요소를 사용하는 경우의 정확성은 훈련 데이터 집합에 대해서는 99.15%, 평가 데이터 집합에 대해서는 96.93%를 각각 보여 주었다. 이로부터 논문이 제안한 방법에 따라 선택된 특징 요소들을 사용하는 경우가 NSL_KDD 데이터가 상정한 41개 특징 요소 모두를 사용하는 경우 보다 DoS 공격을 탐지하는 데 보다 높은 정확성을 보여 줌을 알 수 있다. 또한 사용되는 특징 요소의 개수가 평균 21.0 개로써 전체 특징 요소의 절반에 해당하는데 이러한 이유 때문에 학습 및 평가에 드는 시간이 모든 요소를 사용하는 경우 1272.88초의 학습 시간과 2.19초의 테스트 시간이 요구된 것에 반해 선택된 특징을 사용하는 경우 평균 420.21초의 학습 시간과 1초 미만의 테스트 시간이 소요된다. 한편 가장 높은 정확성을 보여준 특징 조합은 표 3에서 회색 바탕으로 표시된 경우로 25개의 특징 요소로 구성되어 있고 99% 이상의 정확성을 보여 주었다.

V. 결론

본 논문은 NSL_KDD 데이터를 대상으로 서비스 거부 공격을 탐지하기 위한 최적의 특징 조합 선택을 위한 알고리즘을 제시하였다. 제시한 접근 방법은 특징 조합 선택을 최적화 문제로 정의하고 최적 해를 찾기 위해 지역 탐색 알고리즘을 응용하였다. 제시한 지역 탐색 방법의 비용함수는 군집해의 분류 정확성을 사용하였다는 점에 특색이 있다. 한편 제시한 알고리즘에 의해 선택된 특징 조합의 성능을 확인하고자 인공 신경망을 설계하고 41개로 구성된 모든 특징을 사용했을 때와의 정확성을 비교하였다. 실험 결과 특징 요소의 약 절반만을 사용하고도 모든 특징을 사용했을 때보다 높은 정확성을 보여 주었으며 학습 시간과 테스트 시간 모두에서 3배 이상의 효율성을 확인할 수 있었다.

감사의 글

이 논문은 2014년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2014R1A1A4A01008799)

참고문헌

[1] S. Paliwal and R. Gupta, “Denial-of-Servic, Probing & Remote to User (R2L) Attack Detection using Genetic Algorithm,” International Journal of Computer Application, Vol.60, No.19, 2012.

[2] M. Sabhnani and G. Serpen, “Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context,” Proc. of International Conference on Machine Learning: Models, Technologies, and Applications, 23-26 June 2003, Las Vegas, Nevada, USA, 2003, pp. 209-215

[3] KDD Cup 1999. Available on:http://kdd.ics.uci.edu/databases/kddcup99/kddcup9
9.html, 2007.

[4] E. Kim and S. Kim, “Network Anomaly Detection using Hybrid Feature Selection,” 한국정보보호학회 하계정보보호학술대회 논문집, vol. 16, no. 1, pp.649-653, 2006.

[5] H. G. Kayacik, A. N. Zincir-Heywood, and M. I. Heywood, “Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets,” in Thrid Annual Conference on Privacy, Security and Trust, St. Andrews, New Brunswick, Canada, 2005.

[6] A. A. Olusola, A. S. Oladele, and D. O. Abosede, “Analysis of KDD ’99 Intrusion Detection Dataset for Selection of Relevance Features,” in Proc. of the World Congress on Engineering and Computer Science, Vol. 1, 2010.

[7] S. Parazad, E. Saboori, and A. Allahyar, “Fast Feature Reduction in Intrusion Detection Datasets,” in MIPRO, Proceedings of the 35th International Convention, pp.1023-1029, 2012.

[8] A. H. Sung, and S. Mukkamala, “Identifying Important Features for Intrusion Using Support Vector Machines and Neural Networks,” in 2003 Symposium on Applications and the Internet 2003, pp.209-216, 2003.

[9] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set,” Proc. 2009 IEEE Int. Conf. Comput. Intell. Security Defense Appl. CISDA 2009, pp. 53-58.

[10] NSL_KDD data set. Avalilable on: http://nsl.cs.unb.ca/NSL-KDD/