특집논문-08-13-1-10

통계적 수량화 방법을 이용한 효과적인 네트워크 데이터 비교 방법

조 재 익[a], 김 호 인[a], 문 종 섭[a]‡

Effective and Statistical Quantification Model for Network Data Comparing

Jaeik Cho[a], Hoin Kim[a], and Jongsub Moon[a]‡

요 약

네트워크 데이터 분석에 있어서 추정모델이 얼마나 모집단을 대표 하느냐는 반드시 연구 되어야 한다. 본 논문에서는 네트워크 데이터의 각 추출 가능한 표준 정보를 이용하여 현재 공개되어 사용하고 있는 MIT Lincoln Lab의 네트워크 데이터와 모델링 된 KDD CUP 99 데이터를 비교 분석한다. 비교, 분석에 있어서 두 데이터에 공통으로 포함되고 표준 정보인 프로토콜 정보를 이용하여 분석한다. 분석은 통계적 분석 방법인 대응 분석 방법을 이용하여 분석하고, SVD를 이용해 2차원 공간에 표현하며, 가중 유클리드 거리를 이용해 네트워크 데이터를 수량화 하였다.

Abstract

In the field of network data analysis, the research of how much the estimation data reflects the population data is inevitable. This paper compares and analyzes the well known MIT Lincoln Lab network data, which is composed of collectable standard information from the network with the KDD CUP 99 dataset which was composed from the MIT/LL data. For comparison and analysis, the protocol information of both the data was used. Correspondence analysis was used for analysis, SVD was used for 2 dimensional visualization and weigthed euclidean distance was used for network data quantification.

Keyword : Data Set Comparing, Intrusion Detection, Evaluation Data Set, Data Set Composing, Correspondence Analysis

Ⅰ. 서 론

국내외 네트워크에 관한 침입 및 그에 따른 탐지 기술은 기존의 시그너쳐 기반의 탐지 방법을 확장하거나 대상 네트워크에 맞게 조절한 것이 대부분이다. 또한 이러한 시그너쳐 기반의 침입 탐지에 관한 시스템은 지속적으로 공격 시그너쳐를 추가하여야 하고 실제 공격이 발생 하였을 때에 분석하여야 하기 때문에 새로운 공격 추가는 매우 긴 시간적 간격을 가진다. 추후 침입 탐지 시스템의 성능을 개선하고 이상 탐지 등의 효과적인 방법으로의 연구를 위해서는 정의된 네트워크 모델링 데이터 (네트워크 데이터 셋)가 필요하다. 네트워크의 상태나 혹은 데이터의 흐름을 효과적으로 모델링하고 정확하게 모델링 하기 위해서는 검증 기술이 필요하며, 검증은 실제 네트워크 데이터와 모델링 된 네트워크 데이터의 비교 방법을 명확히 하는데 있다.

그러기 위해서 네트워크에서의 두 데이터 집단 간의 비교는 네트워크 데이터의 일반적인 분포 확인과 두 데이터 집간 간의 비교 방법이 동시에 수행 되어야 한다. 대응 분석 방법은 HCI Tulip Goodman이 제안한 방법으로서[1,2,3,4], 두개의 그룹 데이터에 사이의 관계를 비교, 분석할 수 있는 근거를 제시 하며, 가중 유클리디언 방법을 통해 상대적인

a) 고려대학교 정보보호센터
CIST, Korea University
‡ 교신저자 : 문종섭 (jsmoon@korea.ac.kr)

계량 데이터를 제공한다. 또한, 대응 분석 방법은 다차원의 데이터를 분석하여, 2차원의 테이블로 표현할 수 있는데, 이때 Singular Value Decomposition (SVD) 를 사용하여 데이터를 표현한다. 대응 분석에는 Multiple 대응 분석과 Single 대응 분석이 있는데 Multiple 대응 분석은 모든 그룹들을 동시에 분석하는 것이고, Single 대응 분석은 모든 그룹에서 두 그룹씩 추출하여 분석하는 방법이다[5]. 본 논문에서는 통계적 데이터 비교에서 사용되는 Simple 대응 분석을 이용하여 네트워크 데이터에 적용해 보고 현재 가장 많이 사용하고 있는 MIT/LL의 데이터와 이 데이터의 모델링 된 데이터인 KDD CUP' 99의 데이터를 분석해 보고 문제점을 확인하였다.

본 논문의 구성은 2장에서 대응 분석의 기본이론이 되는 정준 상관 분석 이론을 확인하고 본 논문의 실험에 이용된 MIT/LL 의 데이터와 KDD CUP 99' 데이터를 3장에서 설명하며, 4장에서는 대응 분석 방법을 적용한 네트워크 비교 실험과 그 실험 결과에 대해 확인하였다. 또한 5장에서 본 논문의 결론을 맺는다.

II. 데이터 분석 모델

1. 정준상관 분석

Canonical Correlation model (CCM)은 두 개의 연속형 다변량 자료가 짝으로 연결되어 있는 경우에서 출발한다. CCM 은 Yc를 외적 기준으로 하고 이것의 최적 프로젝션에 대해 X 열공간에서 찾는 것이다[6]. 다음과 같이 정식화 할 수 있다.

$$\max_c \|Xb^*\|^2/n' \quad \text{subject to} \quad \|Yc\|^2/n' = 1 \qquad (1)$$

수식(1)과 같이 CCM 은 본 연구에서 두 집단간의 데이터 비교, 두 집단간의 데이터 분석 연구를 위해 사용한 Correspondence Analysis 에서 기반하고 있는 Model 이다[6].

2. 대응분석

Correspondence Analysis는 앞서 설명한 Canonical Correlation model을 이용하여 변수들 (Feature) 의 상관 관계를 수치상의 거리로 변환하는 방법이다[7]. Singular Vector Decomposition을 이용하여 2차원의 Visible 값으로 나타낼 수 있으며, 본 연구에서는 2차원의 Visible 값의 거리를 유클리디안 방법으로 확인하고 해당 거리를 데이터의 비교 척도로 이용하였다.

각 행과 열의 수량인 x_{im}과 y_{jm}사이의 measure of the correlation은

$$\sum_{i=1}^{I}\sum_{j=1}^{J} x_{im} y_{jm} P_{ij} = \lambda_m$$

RC 상관 모델은

$$P_{ij} = P_{i+}P_{+j}(1 + \sum_{m=1}^{M+} \lambda_m x_{im} y_{jm})$$

여기에서

$$1 \le M \le M_1$$

본 논문에서 사용한 대응 분석 방법은 Canonical correlation에서 확장된 방법으로서 model은

$$P_{ij} = P_{i+}P_{+j}(1 + \sum_{m=1}^{M+} x'_{im} y'_{jm}/\lambda_m)$$

여기에서

$$x'_{im} = \lambda_m x_{im} \ , \ y'_{jm} = \lambda_m x_{jm}$$

III. 예 제

1. MIT 네트워크 패킷 데이터

MIT의 네트워크 패킷 데이터는 MIT Lincoln 연구실 (MIT/LL)에서 개발되었다. 침입 탐지 시스템 성능 평가, 네트워크 패킷 분석 등 여러가지 방면으로 사용하기 위해 공개

된 네트워크 패킷 데이터이다. MIT/LL에서 공개되어 사용되는 네트워크 패킷 데이터는 공군망 데이터를 수집하여, 가공, 변조한 데이터이다. 공군 망의 실제 네트워크 패킷 데이터를 수집하고 개인정보 삭제를 위하여 데이터그램을 변조하였으며, 변조된 데이터그램을 갖는 패킷을 다시 생성 하였다[8].

2. KDD Cup 99' 네트워크 데이터 셋

KDD CUP 99' 네트워크 패킷 데이터는 1998년 MIT/LL에서 만들어진 네트워크 패킷을 이용하여 가공된 데이터이다. KDD CUP에서는 가공된 데이터를 이용하여 침입 탐지 시스템 알고리즘의 성능평가에 기준 데이터로 사용하였다. 여러가지 실제 실험된 SMURF1 등과 같은 공격 데이터와 정상적인 패킷 데이터를 41개의 Feature를 이용하여 데이터 셋으로 표현 하였다. 2초의 시간적인 간격으로 패킷을 구분하고 해당 2초 범위의 패킷에서 41개의 피쳐로 표현하였다[9]. 즉, 41개의 피쳐 속성으로 MIT/LL 의 1998년도 패킷 데이터를 모델링한 데이터이다.

Ⅳ. 네트워크 데이터 셋의 비교

1. 데이터 셋

두 개의 네트워크 데이터 그룹인 MIT 데이터와 MIT 데이터를 이용하여 모델링 된 데이터인 KDD CUP 99'의 데이터를 비교, 분석 하였다.

두가지의 데이터에서 첫번째 데이터는 실제 패킷 데이터로서 KDD CUP 99' 보다가 다양한 변수를 갖고 있으나 / 두번째의 KDD CUP 99의 데이터는 첫번째 데이터를 41개의 피쳐로 데이터를 변환한 데이터 셋 / 으로서 동일하게 사용할 수 있는 정보인 프로토콜을 이용하여 두 개의 데이터 그룹을 비교, 분석하였다.

또한 두가지의 총 데이터 수가 다르기 때문에 데이터의 숫자 또한 동일한 기준에서 비교하기 위하여 빈도 분석 결과를 통한 10000개 기준의 스케일링을 하였다. 빈도분석 결과는 아래와 같다.

표 1. MIT LL Data, Protocol frequency analysis result
Table 1. MIT LL Data, Protocol frequency analysis result

MIT Lincoln Lab.

Protocol	Count	Percent
802.1D	0	0
ARP	104315	0.724136512
CDPv2	11263	0.078185779
FRAGMENT	67929	0.471551254
ICMP	9267	0.064329896
ICMPv6	0	0
IGMP	0	0
NOV-802.2	0	0
OTHERS	173529	1.204608011
TCP	12618567	87.5958883
UDP	1420563	9.861300247
Total	14405433	100

표 2. KDD CUP 99' Data, Protocol frequency analysis result
Table 2. KDD CUP 99' Data, Protocol frequency analysis result

KDD CUP 99'

Protocol	Count	Percent
802.1D	0	0
ARP	0	0
CDPv2	0	0
FRAGMENT	0	0
ICMP	2833544	57.8459747
ICMPv6	0	0
IGMP	0	0
NOV-802.2	0	0
OTHERS	0	0
TCP	1870598	38.18771284
UDP	194287	3.966312465
Total	4898429	100

2. 데이터의 스케일링

4.1의 빈도 분석 결과를 이용하여 4.1에서 설명한 패킷 갯수를 동일하게 구성 후 비교하기 위하여 10000개로 스케일링 하였다. 스케일링은 Scalar multiplication 방법으로서[10]

$$Ra' = r(a_1, a_2) = (ra_1, ra_2)$$

표 3. Protocol data scaling result
Table 3. Protocol data scaling result

	802.1d	ARP	CDPv2	FRAGMENT	ICMP	OTHERS	TCP	UDP	TOTAL
MIT/LL	0	72	8	47	6	120	8760	986	9999
KDD	0	0	0	0	5785	0	3819	397	10001

Scaling 결과에서 두가지의 데이터 모두에서 "0" 인 경우에는 결과에서 제외 하였다.

3. 데이터의 비교

앞서 설명된 2.2 의 방법으로 대응 분석을 하였다. 각 프로토콜의 Scaling된 데이터의 발현 횟수를 이용하여 분석 하였다. 분석 방법은 첫째, 전체 프로토콜 발현 패킷 수를 이용한 두가지 데이터 셋의 분석, 둘째, 일반적으로 가장 많이 발생하는 프로토콜 TCP, UDP의 발현 패킷 수를 이용한 두가지 데이터 셋의 분석, 마지막으로 TCP와 UDP 이외의 발현된 프로토콜에 대한 분석, 세가지로 실험 하였다. 특히 마지막 실험 결과에서는 발생의 횟수는 많지 않지만 네트워크 모델링에서 반드시 고려 되어야 할 소규모 발현 프로토콜에 대해서 따로 확인하여 정밀한 모델링 데이터 비교를 하기 위하여 구분하여 실험하였다. 실험의 결과 다음과 같다.

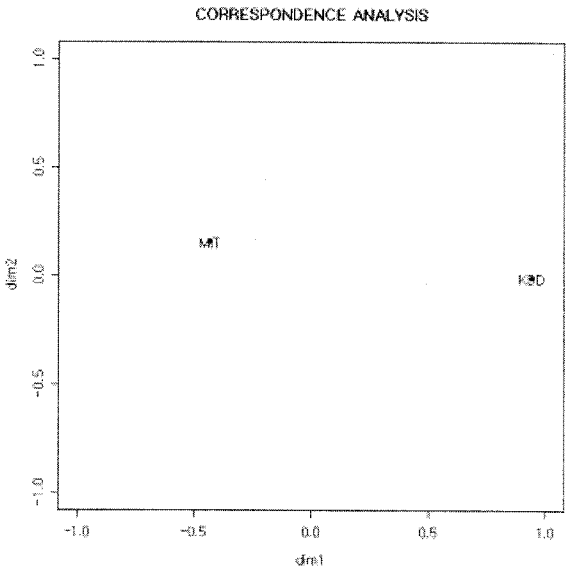


그림 2. TCP, UDP에 수를 이용한 분석
Fig. 2. Frequency analysis using TCP and UDP

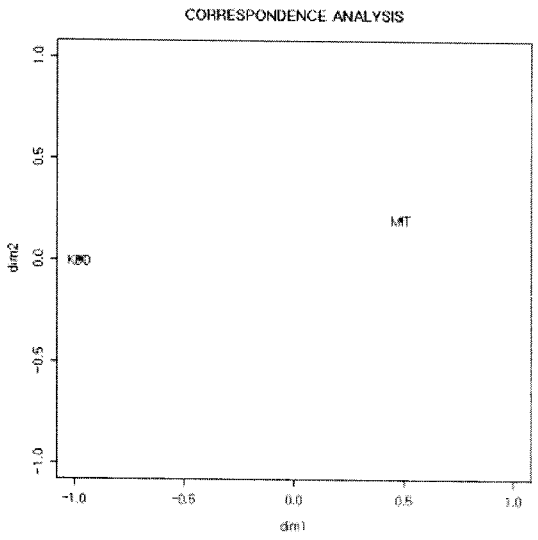


그림 1. 전체 각각의 프로토콜 발현 패킷 수를 이용한 분석
Fig. 1. Packet frequency analysis using all protocols

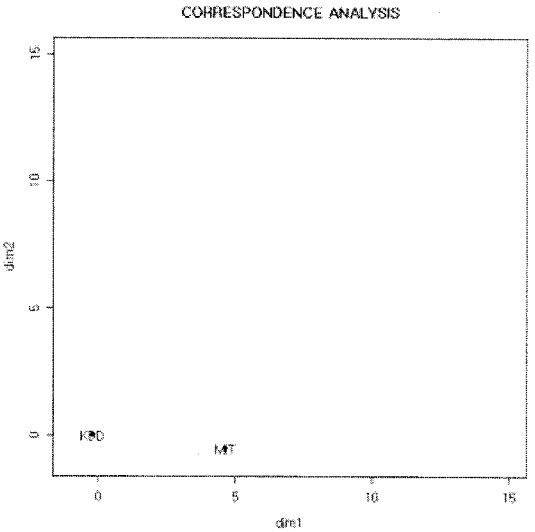


그림 3. TCP, UDP가 아닌 기타 프로토콜 패킷 수를 이용한 분석
Fig. 3. Packet frequency analysis using non-TCP and UDP protocols

그림 1의 두가지 데이터 셋 비교에 대한 수식적 결과는 다음과 같다.

$$x^2_{KDD,MIT} = \sum_{j=1}^{11} \frac{(P_{KDD_j} - P_{MIT_j})^2}{m_j} = 21929.73$$

또한 그림 2의 두 그룹데이터, TCP, UDP, 프로토콜 패킷 대한 결과는 다음과 같다.

$$x^2_{KDD,MIT} = \sum_{j=1}^{11} \frac{(P_{KDD_j} - P_{MIT_j})^2}{m_j} = 19080.72$$

마지막 실험으로 Figure 3의 세 그룹 데이터의 TCP, UDP를 제외한 프로토콜의 패킷 수를 이용한 대응 분석의 수치적 결과는 다음과 같다.

$$x^2_{KDD,MIT} = \sum_{j=1}^{11} \frac{(P_{KDD_j} - P_{MIT_j})^2}{m_j} = 18021.55$$

V. 결 론

네트워크 데이터 분석에 있어서 두 그룹 혹은 여러 그룹에서 동일한 피쳐를 가지고 있는 데이터에 있어서 데이터의 상대적인 비교, 분석에는 대응 분석을 이용하여 상대 적인 관계를 수치화 하여 나타낼 수 있다. 본 논문에서 실험된 내용은 두 데이터 그룹에서 절대적으로 정확한, 혹은 일반화 할 수 없는 데이터를 Correspondence Analysis 를 이용하여 상대 거리를 확인하였다.

참 고 문 헌

[1] Goodman, L.A. Simple models for the analysis of association in cross-classifications having ordered categories. J. Am. Statist. Assoc. 74, 537-552. 1979

[2] Goodman, L.A. Association models and canonical correlation in the analysis of cross-classifications having ordered categories. J. Am. Statist. Assoc. 76, 320-334. 1981

[3] Goodman, L.A. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. Ann. Statist. 13, 10-69. 1985

[4] Goodman, L.A. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables (with discussion). Int. Statist. Rev. 54, 243-309. 1986

[5] MH Huh. Correspondence Analysis of Two-way Contingency Tables with Ordered Column Categories. International Statistical Institute. Vol. 52. Pp59-60. 1999.

[6] James Lattin, J. Douglas Carroll, Paul E. Green. Analyzing Multivariate Data. Thomson. Pp318, 2003

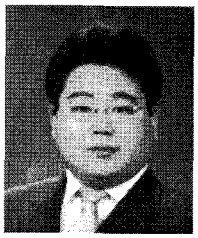
[7] Alan Agresti. Categorical Data Analysis. Pp382. Wiley. 2002

[8] J. W. Haines. 1999 DARPA Intrusion Detection Evaluation. Technical Report 1062. MIT Lincoln Laboratory. 2001

[9] Saharon Rosset, Aron Inger. KDD-cup 99. ACM SIGKDD Explorations Newsletter. KDD-99 Conference report. 2000

[10] James Lattin, J. Douglas Carroll, Paul E. Green. Analyzing Multivariate Data. Thomson. pp25, 2003

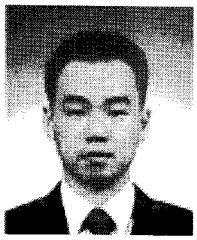
저 자 소 개



조 재 익

- 2005년 : 동국대학교 컴퓨터학 학사
- 2008년 : 고려대학교 정보경영공학전문대학원 석사
- 2008 ~ 현재 : 고려대학교 정보경영공학전문대학원 박사과정
- 주관심분야 : 패턴인식, 네트워크 모델링, 시스템 보안

저 자 소 개



김 호 인

- 2007년 : 동국대학교 멀티미디어공학 학사
- 2007년 ~ 현재 : 고려대학교 정보경영공학전문대학원 석사과정
- 주관심분야 : 네트워크 모델링, 네트워크 보안, 시스템 보안

문 종 섭

- 1991년 : Illinois Institute of Technology 전산학 박사
- 2002년 ~ 현재 : 고려대학교 전자 및 정보공학부 교수
- 2001년 ~ 현재 : 고려대학교 정보경영공학전문대학원 겸임교수
- 주관심분야 : 신경망 이론, 패턴인식, 시스템 보안, 네트워크 보안