이상탐지를 위한 Hybrid Product of Experts 모델과 학습 알고리즘

김권일 [1O] 장병탁 [2]

서울대학교 협동과정 뇌과학전공 [1] 서울대학교 컴퓨터공학부 [2]

{kikim, btzhang}@bi.snu.ac.kr

Hybrid Product of Experts Model and Learning Algorithm for Anomaly Detection

Kwonill Kim[1O] Byoung-Tak Zhang[2]

Interdisciplinary Program in Neuroscience, Seoul National University[1]

School of Computer Science & Engineering, Seoul National University[2]

요 약

이상탐지(anomaly detection)는 기계학습과 데이터마이닝의 중요한 문제로서 여러 응용 분야에서 다양한 형태로 연구되고 있으나, 해결을 어렵게 하는 몇 가지 특징들을 가진다. 이상 데이터의 수가 정상 데이터에 비해 매우 적은 경우가 많고, 정상 패턴과 매우 유사하거나, 알려진 이상 패턴과 다른 새로운 이상 패턴이 발생하기도 한다. 이를 극복하기 위해 정상 데이터를 잘 학습하고, 학습된 정상 데이터의 특징과 구분되는 이상 데이터의 특징을 추출할 수 있는 모델과 학습 알고리즘이 필요하다. 본 논문에서는 이러한 특성을 가지는 새로운 Product of Experts(PoE) 모델과 학습 알고리즘을 제안한다. 다양한 PoE 모델들을 정상/이상 데이터의 특성에 맞게 적용하여 효과적으로 표현하고, 정상 패턴과 큰 차이를 보이는 이상 특성을 효율적으로 학습할 수 있는 방법을 제시하였다. German Credit data에 대해, PoE의 대표적 모델인 Restricted Boltzmann Machine (RBM)을 적용하여 학습한 결과, 일정 수준의 분류성능과 함께, 정상/이상 feature의 발화 패턴이 뚜렷하게 나뉘는 것을 확인할 수 있었다. 앞으로 이 모델을 적용하여 이상감지 문제에 deep network 를 활용하는 연구가 이어질 것으로 기대된다.

1. 서 론

이상탐지(anomaly detection)는 예측된 행동에 따르지 않는 패턴을 데이터에서 찾아내는 문제로서 anomaly, outlier, novelty, discordant, exception, aberration, surprise 등 다양한 표현으로 침입감지(intrusion detection), 사기감지(fraud detection), 보건의료, 센서네트워크와 같은 여러 분야에서 연구되고 있다. 그러나 이상탐지는 기계학습 관점에서 다음과 같은 난점을 가진다.[1][2][3]

- 정상 데이터에 비해 이상 데이터의 수가 매우 부족한 경우가 많다. 이로 인해 이상 패턴 학습과 분류가 매우 어려워진다.
- 이상 패턴의 정의가 분야마다 다르며, 정상 패턴과 이상 패턴을 명확히 구분하기 힘들 수 있다. 사기탐지(fraud detection) 분야에서는 정상 패턴으로 위장하는 경우도 흔하다.
- 알려진 이상패턴들과 다른, 새로운 이상패턴이 등장하기도 한다.

이러한 특징들로 인해 이상탐지 문제에 적합한 기계학습 모델이 필요하며, 정상/이상 데이터의 확률분포를 component들의 합으로 표현하는 mixture 모델이 흔히 사용되고 있다. 다양한 component들로 정상/이상 데이터의 상이한 특성을 효과적으로 표현하고, 정상 모델과 불일치하는 정도를 척도로 삼아 새로운 이상패턴을 감지할 수도 있다.

Product of experts(PoE) 모델은 mixture 모델과 달리 데이터의 확률 분포를 feature 함수들의 곱으로 표현하는 모델로서, Deep Belief Networks(DBN) 의 구성 단위로 사용되는 Restricted Boltzmann Machine(RBM)나 Markov Random Field(MRF)가 여기에 속하며, 다양한 분야에서 좋은 성능을 보이고 있다.[4][5]

본 논문에서는 Product of experts(PoE) 모델을 이상탐지에 적용하여, 정상 데이터의 확률분포와 이상 데이터의 패턴을 효과적으로 표현하는 Hybrid PoE 모델과 이를 효율적으로 학습하는 학습 알고리즘을 제안하고자 한다. 2장에서 Hybrid PoE 모델과 학습 알고리즘을 제안하고, 3장에서 이를 기반으로 anomaly detection을 위한 classifier를 구성하는 방법을 소개한 다음, 4장에서 German Credit dataset을 사용하여 모델과 학습 알고리즘을 검증하도록 하겠다.

2. Hybrid Product of Expert Model

PoE 모델은 확률분포를 feature 함수, $f_m(\mathbf{x};\boldsymbol{\theta}_m)$ 들의 곱으로 아래 식과 같이 표현할 수 있는 모델을 뜻한다.

$$P(\mathbf{x};\Theta)=\frac{1}{Z}P^*(\mathbf{x};\Theta)=\frac{1}{Z}\prod_{m=1}^{M}f_m(\mathbf{x};\boldsymbol{\theta}_m),$$
$$\text{where } \Theta=\{\boldsymbol{\theta}_m\}_{m=1}^{M} \text{ and } Z=\int P^*(\mathbf{x};\Theta)d\mathbf{x}.$$

$\boldsymbol{\theta}_m$ 는 각 feature 함수의 learning parameter를, Z 는 normalize 항, 그리고 $P^*(\mathbf{x};\Theta)$ 는 unnormalized 확률분포 함수를 의미한다.

학습 데이터 $\mathcal{D}=\{\mathbf{x}^{(n)};y^{(n)}\}_{n=1}^{|\mathcal{D}|}$, $y\in\{-1,+1\}$ 가 주어졌을 때, 이를 정상 데이터 $\mathcal{N}=\{\mathbf{x}^{(n)}\mid y^{(n)}=-1,\ 1\le n\le|\mathcal{D}|\}$ 와 이상 데이터 $\mathcal{A}=\{\mathbf{x}^{(n)}\mid y^{(n)}=+1,\ 1\le n\le|\mathcal{D}|\}$ 로 나누고, 각각 확률분포 $P_{\mathcal{N}}(\mathbf{x})$ 와 $P_{\mathcal{A}}(\mathbf{x})$ 를 따른다고 가정하자. 그리고 정상 데이터 $\mathcal{N}$ 을 PoE 모델 $P(\mathbf{x};\Theta_{\mathcal{N}})$ 로 학습하고, 이 정상 모델 $P(\mathbf{x};\Theta_{\mathcal{N}})$ 에 비정상 feature 함수를 추가한 모델 $P(\mathbf{x};\Theta_{\mathcal{N}},\Theta_{\mathcal{A}})$ 로 이상 데이터 $\mathcal{A}$ 를 학습하려 한다. 이는 아래 likelihood를 최대화하는 learning parameter $\Theta_{\mathcal{N}},\Theta_{\mathcal{A}}$ 를 구하는 문제로 정의 되며,

$$L(\Theta_{\mathcal{N}},\Theta_{\mathcal{A}})=\frac{1}{|\mathcal{N}|}\sum_{\mathbf{x}\in\mathcal{N}}\ln P(\mathbf{x};\Theta_{\mathcal{N}})+\frac{1}{|\mathcal{A}|}\sum_{\mathbf{x}\in\mathcal{A}}\ln P(\mathbf{x};\Theta_{\mathcal{N}},\Theta_{\mathcal{A}})$$
$$=L(\Theta_{\mathcal{A}})+L(\Theta_{\mathcal{A}}\mid\Theta_{\mathcal{N}})$$

용이한 계산을 위해 이를 두 항으로 나누어 다음 알고리즘으로 학습하였다.

Alg. 1. $\Theta_{\mathcal{N}}^*=\arg\max_{\Theta_{\mathcal{N}}}L(\Theta_{\mathcal{N}})$

$$\theta_{m,\text{new}}=\theta_{m,\text{old}}+\partial L(\Theta_{\mathcal{N}})/\partial\theta_m$$
$$\frac{\partial L(\Theta_{\mathcal{N}})}{\partial\theta_{m_{\mathcal{N}}}}=\frac{1}{|\mathcal{N}|}\sum_{\mathbf{x}\in\mathcal{N}}\frac{\partial\ln f_{m_{\mathcal{N}}}(\mathbf{x})}{\partial\theta_{m_{\mathcal{N}}}}-\mathbb{E}\left[\frac{\partial\ln f_{m_{\mathcal{N}}}(\mathbf{X})}{\partial\theta_{m_{\mathcal{N}}}}\right]_{P(\mathbf{X};\Theta_{\mathcal{N}})}$$

Alg. 2. $\Theta_{\mathcal{A}}^*=\arg\max_{\Theta_{\mathcal{A}}}L(\Theta_{\mathcal{A}}\mid\Theta_{\mathcal{N}}^*)$

$$\frac{\partial L(\Theta_{\mathcal{A}}\mid\Theta_{\mathcal{N}}^*)}{\partial\theta_{m_{\mathcal{A}}}}=\frac{1}{|\mathcal{A}|}\sum_{\mathbf{x}\in\mathcal{A}}\frac{\partial\ln f_{m_{\mathcal{A}}}(\mathbf{x})}{\partial\theta_{m_{\mathcal{A}}}}-\mathbb{E}\left[\frac{\partial\ln f_{m_{\mathcal{A}}}(\mathbf{X})}{\partial\theta_{m_{\mathcal{A}}}}\right]_{P(\mathbf{X};\Theta_{\mathcal{N}}^*,\Theta_{\mathcal{A}})}$$

그런데, $P(\mathbf{x};\Theta_{\mathcal{N}},\Theta_{\mathcal{A}})$ 는 PoE모델이므로 다음과 같이 정상 패턴과 비정상 패턴을 표현하는 함수들의 곱으로 분리할 수 있다.

$$P(\mathbf{x};\Theta_{\mathcal{N}},\Theta_{\mathcal{A}})=\frac{1}{Z}\prod_{m_{\mathcal{N}}}^{M_{\mathcal{N}}}f_{m_{\mathcal{N}}}(\mathbf{x};\boldsymbol{\theta}_{m_{\mathcal{N}}})\prod_{m_{\mathcal{A}}}^{M_{\mathcal{A}}}f_{m_{\mathcal{A}}}(\mathbf{x};\boldsymbol{\theta}_{m_{\mathcal{A}}})$$
$$=\frac{P_{\mathcal{N}}^*(\mathbf{x};\Theta_{\mathcal{N}})P_{\mathcal{A}}^*(\mathbf{x};\Theta_{\mathcal{A}})}{\int P_{\mathcal{N}}^*(\mathbf{x};\Theta_{\mathcal{N}})P_{\mathcal{A}}^*(\mathbf{x};\Theta_{\mathcal{A}})d\mathbf{x}}$$

이를 이용해 $L(\Theta_{\mathcal{A}}\mid\Theta_{\mathcal{N}}^*)$ 를 정리하면 다음과 같고,

$$L(\Theta_{\mathcal{A}}\mid\Theta_{\mathcal{N}}^*)=\frac{1}{|\mathcal{A}|}\sum_{\mathbf{x}\in\mathcal{A}}\left\{\ln P(\mathbf{x};\Theta_{\mathcal{N}}^*)+\ln P(\mathbf{x};\Theta_{\mathcal{A}})\right\}$$
$$-\ln\sum_{\mathbf{x}}P(\mathbf{x};\Theta_{\mathcal{N}}^*)P(\mathbf{x};\Theta_{\mathcal{A}})$$

여기에 Jensen's inequality를 적용하면 다음과 같이 upper bound를 구할 수 있다.

$$L(\Theta_{\mathcal{A}}\mid\Theta_{\mathcal{N}}^*)\le\frac{1}{|\mathcal{A}|}\sum_{\mathbf{x}\in\mathcal{A}}\left\{\ln P(\mathbf{x};\Theta_{\mathcal{N}}^*)+\ln P(\mathbf{x};\Theta_{\mathcal{A}})\right\}$$
$$+D_{KL}(P(\mathbf{x};\Theta_{\mathcal{N}}^*)\,\|\,P(\mathbf{x};\Theta_{\mathcal{A}}))$$

위 식에서 두 번째 항은 Alg.1에서 학습된 정상모델 $P(\mathbf{x};\Theta_{\mathcal{N}}^*)$ 과 이상데이터 모델링을 위해 추가된 부분 $P(\mathbf{x};\Theta_{\mathcal{A}})$ 사이의 KL divergence 이며, 이는 Alg.2가 둘 사이의 차이를 증가시키는 방향으로 $\Theta_{\mathcal{A}}$ 를 변경함을 뜻한다. 따라서, hybrid PoE 모델은 Alg.1을 통해 정상 데이터를 표현하는 feature들을 추출하고, Alg.2에서는 정상 데이터 모델과 불일치하는 이상 패턴의 feature를 추출하게 된다.

3. Hybrid PoE Classifier

학습된 정상/이상 모델에 output layer를 추가하여 이상감지를 위한 classifier를 구현할 수 있다. PoE 모델로 RBM을 사용한다면, Fig.1 과 같은 구조가 된다. 이는 hybrid PoE 모델 학습을 통해 추출된 feature들을 사용해 classifier를 구축하는 것으로, 보다 분류 성능의 향상을 기대할 수 있다.

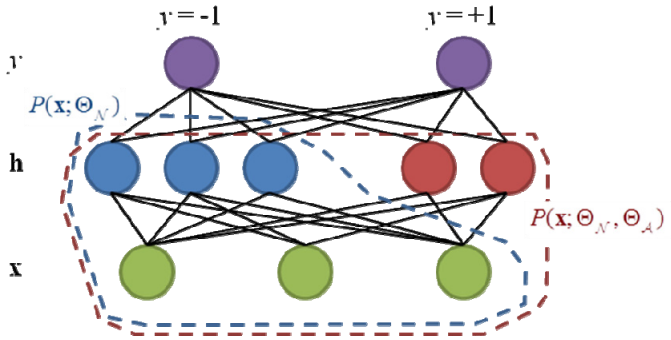

Fig. 1. RBM을 사용한 hybrid PoE classifier

4. 실험 및 결과

본 논문에서 제시한 모델과 학습 알고리즘을 검증하기 위하여 Australian Credit dataset[6][7]을 binary로 변환하여 실험하였다. RBM을 PoE 모델로 사용한 Hybrid PoE 모델을 Matlab 2013a로 구현하였으며, 다른 비교 실험들은 Weka 3.7.5를 이용하였다. 모든 실험은 3-fold cross-validation을 5회 반복하여 실행하였고, accuracy와 AUC(Area under Curve)를 비교하였다. 그 결과, Hybrid PoE 모델은 좋은 분류 성능을 보였다.(Table 1)

Table 1. Accuracy and AUC(Area under Curve). HPoE: hybrid PoE, NB: Naive Bayes, k-NN: k=5, DT: Decision Tree (C4.5), SVM: RBM kernel

	Accuracy	AUC
HPoE	**0.876** ± 0.01	0.921 ± 0.002
NB	0.868 ± 0.003	0.920 ± 0.002
DT	0.862 ± 0.005	0.869 ± 0.007
k-NN	0.866 ± 0.007	**0.926** ± 0.005
SVM	**0.877** ± 0.004	0.875 ± 0.004

또한, hidden layer의 발화 패턴을 관찰하였을 때, Fig. 2. 와 같이 정상/이상 데이터와 정상/이상 feature들 사이에 뚜렷한 상관관계가 나타남을 확인할 수 있다. 이는 Hybrid PoE 모델을 이용해 정상 패턴과 구분되는 이상 패턴을 추출할 수 있는 가능성을 보여준다.

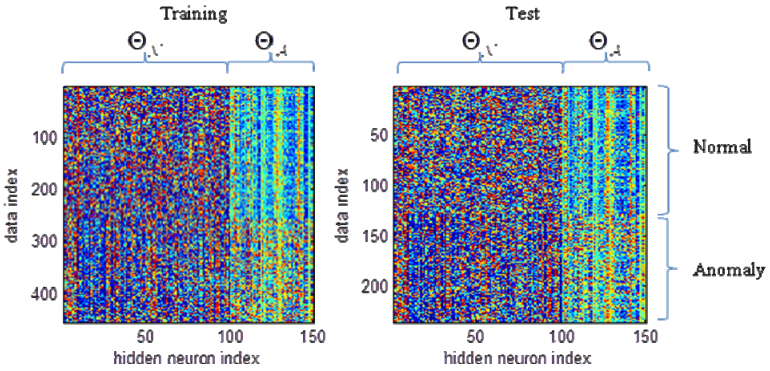

Fig. 2. Hidden layer의 발화 패턴

5. 결론

본 논문에서는 Product of experts(PoE) 모델을 이상탐지에 적용하여, 정상 데이터의 확률분포와 이상 데이터의 패턴을 효과적으로 표현하는 Hybrid PoE 모델과, 이를 효율적으로 학습하는 학습 알고리즘을 제안하였다. 그리고 대표적인 PoE 모델인 RBM에 적용하고 classifier로 구성하여, 좋은 분류 성능과 정상/이상 패턴 추출이 가능함을 보였다. 앞으로 이 모델을 deep network에 적용하는 방안과 해석 가능한 정상/이상 패턴을 추출할 수 있는 PoE 모델에 대한 연구가 이루어질 것으로 기대된다.

Acknowledgement

이 논문은 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2010-0017734-Videome, NRF-2013M3B5A2035921- HyperIntelligence), 정부(산업통상자원부)의 재원으로 한국산업기술평가관리원 지원(KEIT-10035348-mLife, KEIT-10044009)을 일부 받았음.

참고문헌

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” ACM Computing Surveys (CSUR), vol. 41, no. 3, pp. 15:1–15:58, 2009.

[2] G. Blanchard, G. Lee, and C. Scott, “Semi-supervised novelty detection,” The Journal of Machine Learning Research, vol. 11, pp. 2973–3009, 2010.

[3] G. Weiss, “Mining with rarity: a unifying framework,” ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 7–19, 2004.

[4] Y. Bengio, “Learning Deep Architectures for AI,” Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1−127, 2009.

[5] M. Welling, "Product of experts," Scholarpedia, 2(10):3879, 2007.

[6] L. Zhou, K. Lai, and J. Yen, “Credit scoring models with AUC maximization based on weighted SVM,” International Journal of Information Technology & Decision Making, vol. 8, no. 4, pp. 677−696, 2009.

[7] K. Bache, and M. Lichman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science , 2013.