네트워크 AI 분야에 적용한 데이터 라벨링 자동화에 관한 실증 연구

이민호, 고승현, 이종필, 최아솔, 김태영
KT 융합기술원 Infra 연구소

{min.ho.lee, koyosae, jongpil.lee, a-sol.choi , kim.taeyoung}@kt.com

An Empirical Study on Automated Labeling System for Network Infra Data

Lee Minho, Koh Seoung Hyun, Lee Jong Pil, A-sol Choi, Kim Tae Young
KT ICT Infra R&D Lab.

요 약

본 논문에서는 복잡하고 고객마다 다양한 네트워크 장비에서 발생하는 데이터를 그 특성과 용도에 따라 라벨링 처리과정을 자동화하여, 네트워크 관제를 위한 AI 모델 훈련을 위한 학습용 데이터 형태로 가공하는 방법에 대한 실증 연구 사례를 다루고 있다. 자동화된 라벨링 시스템을 통해 AI 기반 네트워크 관제시스템의 빠른 최적화와 네트워크 장비 및 망 구조의 변화에도 효과적으로 대응하여 실시간 유지보수를 구현하는 것에 기여하고자 한다.

Ⅰ. 서 론

AI 기반 네트워크 관제 지능화에 대한 시장의 기대가 증가하는 상황에서, 통신 사업자가 고객 니즈에 따라 다양한 AI 관제 서비스를 빠르게 구축해 제공하는 것은 중요한 역량으로 평가될 것이다. 또한 AI 관제 서비스는 점점 복잡해지고 다양해지는 통신 환경에 대해 능동적으로 대응할 수 있는 기능을 갖추어야 한다. 이에 대해 본 논문에서는 네트워크 AI 개발과 유지보수 가속화를 위한 자가학습 체계 마련을 위해, 네트워크 각 도메인에서 발생하는 데이터를 특성과 용도에 따라 라벨링 처리하여 AI 용 학습데이터의 형태로 가공하는 방법에 대해 기술한다. 실증 연구는 데이터의 피쳐화를 자동으로 인지하여 라벨링 기준을 선정하는 방법과[1], 사용자 지식 기반의 로직 적용을 통해 라벨링 기준을 최적화하는 방법[2,3,4]을 활용하여 이루어졌다. 이를 통해, 기존 수작업을 통해 이루어지던 AI 학습데이터의 라벨링 과정을 자동화하여 네트워크 관제 AI 의 최적화 작업을 앞당기는 것을 목적으로 한다.

Ⅱ. 본론

통신 서비스를 제공하기 위해 일반적으로 선로, 전송, IP, 무선 등 Multi-Layer 기반의 통신망이 구축되어 통신사업자에 의해 관리되며, 각 도메인 별로 안정적인 운용을 위해 감시되는 데이터는 로그, 경보, 전표부터 트래픽과 성능 통계 등 필요에 따라 다양하게 구성되어 있다. AI 기반의 네트워크 관제 지능화를 위해서는 각 도메인에 최적화된 데이터의 활용과 이를 처리하기 위한 AI 모델의 적용이 수반되어야 하며, 이 때 데이터를 AI 모델에 학습시켜 운용하기 위해서는 각 데이터의 특성과 모델에 적합한 형태로의 데이터 가공 과정이 이루어지게 된다.

하지만 AI 기반 관제시스템을 구축하는 과정에 있어서 전체 업무 중 데이터 가공이 차지하는 비중은 약 80%로, 인력·시간·비용적 측면에서 많은 투자를 요구하게 된다. 특히 25%의 업무 비중이 효과적으로 AI 모델을 학습시키기 위한 데이터 라벨링 업무에 집중되어 있어, 이를 자동화 시키는 것은 전체 개발 과정에 있어서 효과적인 자원 절감 효과로 이어지게 된다.

본 논문에서는 데이터의 특성과 용도에 따른 데이터 라벨링 자동화 시스템을 구현하기 위해, 다양한 도메인 상에서 현재 실증되어 운용중인 AI 기반 관제 시스템들의 데이터 활용 사례를 기반으로 설계를 진행하였다. 이를 통해 네트워크에서 발생되는 데이터를 경보·로그·전표 등의 Text 형 타입과, 트래픽·성능지표·통계정보 등의 Time-Series 형 타입으로 분류하였으며, 이를 통해 도메인 별 운용 환경에 맞는 지도·비지도 학습 AI 모델에 적합한 형태로 라벨링 작업을 수행할 수 있는 형태로 구성되었다. 그림 1 은 학습 목적에 따른, 데이터 형태 별 라벨링 과정에 대한 흐름도를 나타낸다.

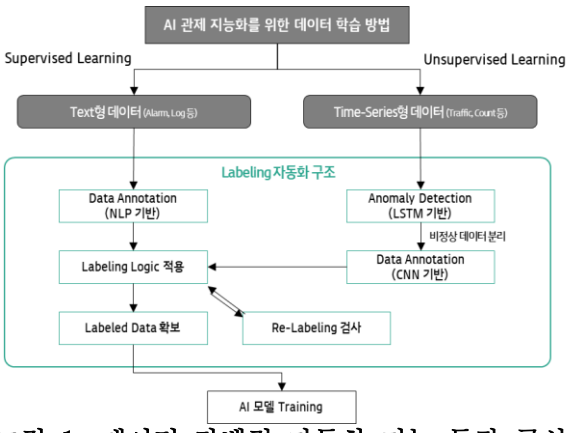

그림 1. 데이터 라벨링 자동화 기능 동작 구성

Text 형질의 데이터는 망 구성 및 통신 장비의 장애 발생 현황과 연계되어 지도학습에 곧바로 활용되는 사례가 많지만, Time-Series 형질의 데이터는 일반적으로 통신서비스의 제공 현황에 밀접한 지표로 활용되기에 이중화 구성과 같은 보완책으로 인해 장비 장애가 서비스 제공에 직결적인 영향을 끼치는 경우가

적은 만큼 장애에 관련된 데이터의 확보가 어렵고, 이로 인해 지도학습을 곧바로 수행하기 어려운 현상이 발생한다. 따라서 Time-Series 형 데이터에 대해서는 우선 LSTM 등의 모델을 활용해 정상/비정상 데이터를 분류하는 비지도학습 형식의 학습 방법이 주로 적용되며, 분류한 비정상 데이터에 대해 지도학습을 적용하는 방법으로의 접근 방식을 고려하여야 한다.

데이터 라벨링 자동화를 위해서는 데이터의 분석과 분류 과정을 거치게 되며, 자동으로 분류된 데이터에 대한 커스터마이즈 과정이 고려되어야 한다. 이를 위해 본 연구에서는 데이터 라벨링을 구현하기 위한 각 기술을 3 가지 단계로 구분하여 구성하였다.

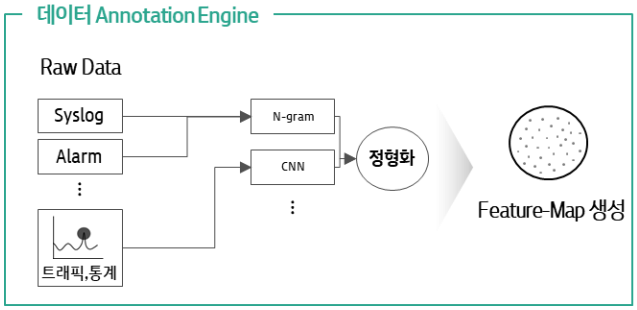

그림 2. Text/Time-Series Data Annotation 구성

첫번째는 데이터에서 Annotation Feature 를 추출하기 위한 단계이다. Text 데이터는 N-gram 을 비롯한 NLP 기법을 통해 특성 정보로 변환되며, Time-Series 형 데이터는 비정상적인 데이터 패턴으로 구분된 영역에 한해 CNN 과 같은 방법을 통해 그 특성 정보가 추출된다. 이를 기반으로 유사 특성을 가진 데이터가 밀집되어 표현될 수 있는 벡터 플레인 상에 데이터를 분포시킬 수 있다. 그림 2 는 이와 같이 네트워크 각 도메인 데이터에 대해 Feature-Map 을 생성하는 단계에 대해 표현하고 있다.

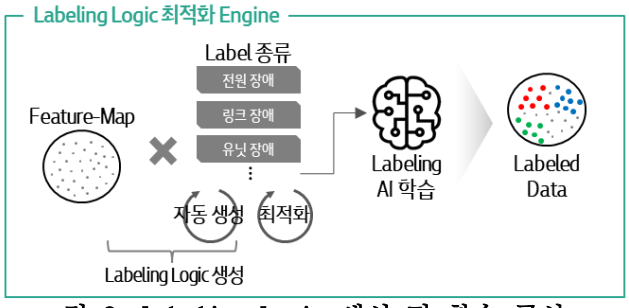

그림 3. Labeling Logic 생성 및 학습 구성

두번째 단계는 Labeling Logic 을 구성하여 Feature 화 된 데이터를 대상으로 실질적인 라벨링 정보를 지정하는 것이다. Labeling Logic 은 두 가지 방법으로 생성되는데, 기존에 트레이닝된 분류 모델을 통한 1 차 Logic 생성과 이를 실제 AI 모델 운용 현황에 따라 편집·추가하는 2 차 Logic 생성의 기능을 제공한다. 1 차 Logic 생성은 Decision Tree 나 KNN 과 같은 방법으로 데이터를 Feature 유사도에 따라 구분하는 방법으로 적용되며, 실제 AI 관제에서는 특이한 케이스에 발생되는 데이터에 대한 관측이 필요한 경우가 존재하므로 이를 해소하기 위해 템플릿 형태로 제공되는 Logic 편집 기능을 통해 필요한 데이터 라벨링의 분류 기준을 추가할 수 있도록 설계되었다. 그림 3 은 Labeling Logic 을 Feature-Map 에 적용하여 Labeling 을 위한 AI 모델을 학습시키는 과정을 보여준다.

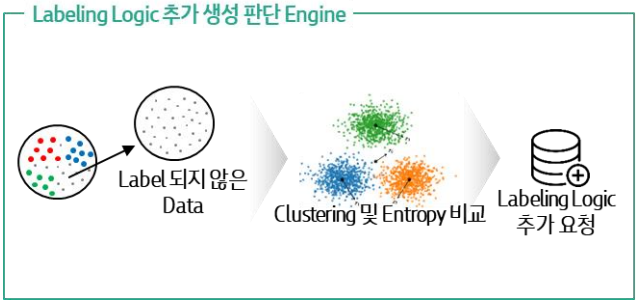

그림 4. Re-Labeling 처리 단계 구성

마지막 단계는 앞선 Labeling Logic 에 의해 1 차 라벨링 처리된 데이터 중 Logic 의 커버리지에 포함되지 못해 라벨링이 되지 않은 데이터에 대한 Re-Labeling 처리의 과정이다. 라벨링되지 않은 데이터 중 라벨링이 반드시 되어야 할 유의미한 데이터 군집을 추출하는 단계로, 본 단계에서는 데이터를 분포도와 군집성에 대한 분석 결과를 기반으로 하여 가중치를 계산하고 이를 기반으로 Labeling Logic 의 추가를 통해 AI 모델 학습에 필요한 데이터가 충분히 라벨링 가공이 완료될 수 있도록 지원한다. 그림 4 는 라벨링 처리가 되지 않은 데이터에 대한 분석을 통해 사용자에게 Re-Labeling 을 위한 Labeling Logic 추가를 요청하는 과정을 표시한다.

위 3 단계를 통한 방법으로 네트워크 각 도메인에서 발생되는 데이터에 대한 자동화된 라벨링 가공을 수행할 수 있으며, Labeling Logic 의 편집 과정을 통해 특이 조건에 대한 라벨링 적용 기준의 유지보수가 가능한 형태로의 라벨링 자동화 시스템 구축이 가능하다. 이를 통해 지속적인 학습데이터 확보가 가능해지며, 기 구축된 Labeling Logic 을 활용하여 신규 장비에 대한 AI 모델의 수용 속도도 상승시키는 효과를 기대할 수 있다.

Ⅲ. 결론

본 논문에서는 자동화된 라벨링 기법을 적용하여 네트워크의 여러 도메인에서 발생하는 데이터를 학습데이터 형태로 빠르게 가공하는 방법을 제안하였다. 기존에 수작업으로 이루어지던 데이터 가공 과정을 자동화 시킴으로써, 수작업 대비 수백배 향상된 속도로 학습데이터를 생성함으로써 AI 모델 학습의 단계로의 진입 속도를 가속화하는 것이 본 논문이 기여하고자 하는 부분이다. 이를 위해 본 논문에서는 Text 와 Time-Series 형태로 데이터를 나누어 처리하고, 이 데이터에 대한 Annotation 과 Labeling Logic 적용, 그리고 Re-Labeling 과정을 자동화하는 방법에 대하여 기술하였다. 이러한 방식을 통해 통신사업자는 보다 빠르게 AI 관제 업무를 지능화하는데 활용할 수 있고, 고객 장비에 대한 수용을 가속화하여 AI 서비스의 제공에 필요한 시간을 단축할 수 있을 것이다. 향후 연구는 각 도메인 별 발생 데이터를 보다 세분화하여 라벨링을 위한 로직 구성을 보다 심층적으로 구성해 제공화하는 것과, AI 모델 학습과 제공의 단계까지도 자동화하여 AI 기반 관제 지능화의 전체 과정을 자동화할 수 있는 방안에 대한 고민이 필요할 것이다.

참 고 문 헌

[1] Paroma V, " Snuba: automating weak supervision to label training data." Proceedings of the VLDB Endowment, pp. 709-730, 2020.

[2] Alexander R."Snorkel: Rapid Training Data Creation with Weak Supervision." Proceedings of the VLDB Endowment, pp. 269-282, 2017.

[3] Alexander R. "Data programming: Creating large training sets, quickly." In Advances in Neural Information Processing Systems, pp. 3567-3575, 2016.

[4] P. Varma. “Inferring generative model structure with static analysis.” In Advances in Neural Information Processing Systems, pp. 239-249, 2017