인텔리전트 데이터베이스 시스템에서 자율 기계 학습을 위한 점진적 학습 기법

김윤아, 임종태, 함동호, 김남영, 이소민, 신보경, 이현병, 유재수*

충북대학교

rud5356@naver.com, jtlim@cbnu.ac.kr, ii8858@naver.com, minstrel68@naver.com,
{somin, sbk02, lhb, yjs}@cbnu.ac.kr

Incremental Learning Scheme for Autonomous Machine Learning in Intelligent Database Systems

Yuna Kim, Jongtae Lim, Dongho Ham, Nameyoung Kim, Somin Lee, Bokyoung Shin, Hyeonbyeong Lee, Jaesoo Yoo

Chungbuk National University

요 약

최근 빅데이터의 발전과 함께 기계 학습이 중요하게 연구되고 있다. 특히, 지속적으로 생성되는 데이터를 대상으로 한 기계 학습 응용이 요구되고 있다. 기존의 연속 기계 학습 기법들은 새로운 데이터가 수신되면 전체 데이터를 다시 학습하여 생성된 기계 학습 모델의 성능을 일정 수준 이상으로 유지한다. 하지만 학습대상 데이터가 매우 빠르게 축적되는 환경에서 매번 전체 데이터를 사용하여 학습을 수행하고 모델을 생성하는 것은 비효율적일 뿐만 아니라 항상 성능이 향상한다는 것을 보장하지 못한다. 본 논문에서는 인텔리전트 데이터베이스 시스템 환경에서 자율기계 학습을 위한 점진적 학습기법을 제안한다. 제안하는 기법은 점진적 학습을 위하여 학습대상 데이터를 이미 학습에 사용된 데이터와 그렇지 않은 데이터로 구분한다. 그렇게 구분된 학습대상 데이터를 기반으로 새로운 데이터만을 활용한 학습 모델과 점진적 학습을 수행한 학습 모델을 도출한다. 제안하는 기법은 도출된 두개 모델과 기존 학습 모델의 성능을 평가하여 가장 성능이 뛰어난 모델을 서비스에 활용한다.

Ⅰ. 서 론

4차 산업 혁명과 함께 처리해야 할 데이터가 기하급수적으로 증가했다. 특히 기계 학습을 수행함에 있어서 학습대상 데이터가 증가하면 일반적으로 정확도가 향상하는 경향을 가지지만 그에 따른 학습의 복잡도와 처리 시간도 함께 증가하여 높은 성능의 시스템이 요구된다. 따라서 일정시간마다 축적된 모든 데이터를 활용하여 모델을 학습하는 대신 기존 학습된 모델과 새롭게 축적된 데이터를 활용하여 빠르게 학습 모델을 생성하는 점진적 학습 기법이 연구되었다[1, 2]. 하지만 기존 점진적 기계 학습 기법은 학습을 통해 새롭게 생성한 모델은 반드시 성능이 향상하는 것을 가정하여 점진적 학습을 수행한다. 기존 점진적 기계 학습 기법이 수행되는 환경에서는 기계 학습 개발자가 점진적 기계학습을 수행할 때 특정한 기계 학습 알고리즘 수행을 위하여 가공 및 전처리를 수행한 정제된 데이터를 사용하기 때문에 학습대상 데이터가 증가하면 기계 학습 모델의 정확도도 향상하는 경향을 보인다. 하지만 실제 응용(서비스)에 활용되는 데이터의 경우, 값이 누락되어 있거나 적절하지 않은 데이터가 포함되어 있으며, 이런 학습대상 데이터가 가공 및 전처리되지 않은 상태로 학습에 활용되는 인텔리전트 데이터베이스 시스템 환경에서의 자율 기계 학습을 활용하는 응용에서는 시스템이 각 알고리즘에 적합하게 스스로 가공 및 전처리를 하는 것에 제한이 존재하기 때문에 학습대상 데이터가 증가하더라도 정확도가 일정 수준에 머무르거나 오히려 정확도가 감소하는 경우도 발생한다. 본 논문에서는 인텔리전트 데이터베이스 시스템 환경에서 자율 기계 학습을 위한 점진적 학습 기법을 제안한다. 제안하는 기법은 점진적 학습을 위하여 학습대상 데이터를 이미 학습에 사용된 데이터와 그렇지 않은 데이터로 구분한다. 이와 같이 구분된 학습대상 데이터를 기반으로 새롭게 축적된 데이터만을 활용한 학습 모델과 점진적 학습을 수행한 학습 모델을 도출한다. 마지막으로 제안하는 기법은 점진적 학습을 수행한 모델들을 다양하게 결합하여 성능이 뛰어난 학습 모델을 응용에 활용한다.

*교신저자 : yjs@chungbuk.ac.kr

Ⅱ. 제안하는 자율 기계 학습을 위한 점진적 학습 기법

1. 특징

제안하는 점진적 학습 기법은 인텔리전트 데이터베이스 시스템 환경에서 자율 기계 학습을 활용하는 응용을 대상으로 하고 있다. 인텔리전트 데이터베이스 시스템은 학습대상 데이터가 데이터베이스 시스템에 지속적으로 축적되고, 축적된 학습대상 데이터를 인텔리전트 데이터베이스 시스템이 자율적으로 다양한 기계 학습 알고리즘을 적용하여 스스로 응용에 활용 가능한 학습 모델을 생성하는 시스템이다. 이러한 환경에서 기계 학습 개발자는 일정 수준의 기계 학습 알고리즘에 대한 학습대상 데이터의 가공 및 전처리는 가능하지만 모든 기계 학습 알고리즘에 적합한 데이터 가공 및 전처리는 불가능하다. 본 논문에서는 이러한 환경을 고려한 점진적 학습 기법을 제안한다.

2. 점진적 학습을 위한 데이터 이력 관리

점진적 학습을 위해서는 먼저 학습대상 데이터를 이전에 학습에 사용된 데이터와 새롭게 축적된 데이터로 구분해야 한다. 본 논문에서는 데이터의 이력관리를 위하여 PROV 모델을 사용한다.. PROV 모델은 문서의 이력을 관리하기 위한 표준화 모델 중 하나이다. PROV 모델의 클래스로는 Agent, Entity, Activity가 존재하며 Agent는 해당 데이터를 활용하는 개인 및 기관이다. Entity는 생성되어지고 활용되어지는 데이터 및 참조되는 정보이다. 그리고 Activity는 Insert, Delete, Update, Integration과 같은 학습대상 데이터에 대한 작업 수행 내용이다.

그림 1은 학습대상 데이터의 이력 정보 관리 방법과 학습대상 데이터의 구분 방법을 보여준다. 이력 관리는 PROV 모델에 기반 하여 시간, 작업정

보, 관련 데이터(속성)를 유지한다. 이력 정보는 각 관련된 데이터 그룹별로 유지되며 해당 그룹에서 수행된 작업들이 저장된다. 이때 원활한 학습 대상 데이터 이력의 추적을 위하여 Updata는 Delete + Insert로 이력을 기록한다. 인텔리전트 데이터베이스 시스템에서 각각의 학습 모델은 응용에 활용할 목적으로 다양한 데이터를 유지하는데 그 중에 시간과 학습에 사용된 관련 데이터를 포함하고 있다. 점진적 학습을 수행하기 위하여 제안하는 기법은 모델의 생성 시점을 기준으로 이후에 발생한 작업들에 대한 이력 정보를 참조하여 최신 학습대상 데이터로부터 모델이 생성되었던 시점의 학습대상 데이터를 역추적하며 학습에 사용된 학습대상 데이터와 새롭게 추가된 학습대상 데이터로 구분한다. 이력 정보를 기준으로 학습에 사용된 학습대상 데이터와 새롭게 추가된 학습대상 데이터를 구분할 수 있다.

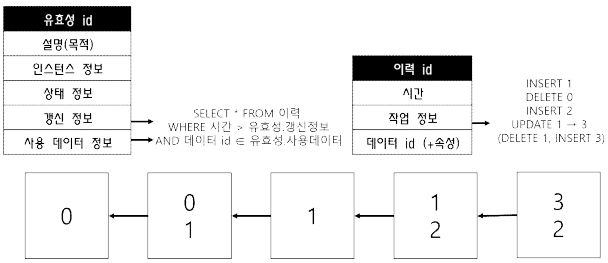

그림 1. 이력 정보 및 학습대상 데이터 구분 방법

3. 모델에 따른 점진적 학습 방법

인텔리전트 데이터베이스에는 다양한 기계 학습 알고리즘에 입력될 수 있는 데이터가 지속적으로 수집된다. 학습대상 데이터가 축적되면 데이터의 특성이 변화함에 따라 기존에 생성되어 활용되고 있던 학습 모델의 성능이 저하될 수 있다. 그래서 인텔리전트 데이터베이스 시스템에서는 응용에서 활용되고 있는 학습 모델에 대하여 유효성 정보를 유지한다. 그림 2는 유효성 정의 및 유효성에 따른 점진적 학습 과정을 보여준다. 유효성은 크게 데이터의 양, 시간, 요구되는 성능 등으로 정의할 수 있다. 데이터의 양에 따른 유효성은 마지막 학습 모델을 생성한 시점의 데이터양을 기준으로 사용자가 정의한 일정량 이상의 데이터가 추가되면 모델의 성능을 향상시키기 위한 점진적 학습을 수행하는 개념이다. 시간에 따른 유효성은 마지막 학습 모델을 생성한 시간을 기준으로 사용자가 정의한 일정 시간 이상이 지나면 모델의 성능을 향상시키기 위한 점진적 학습을 수행하는 개념이다. 마지막으로 요구되는 성능에 따른 유효성은 학습 모델이 활용될 때마다 성능을 평가하여 사용자가 결정한 임계치 이상의 성능을 만족하지 않을 경우 성능을 향상시키기 위한 점진적 학습을 수행한다.

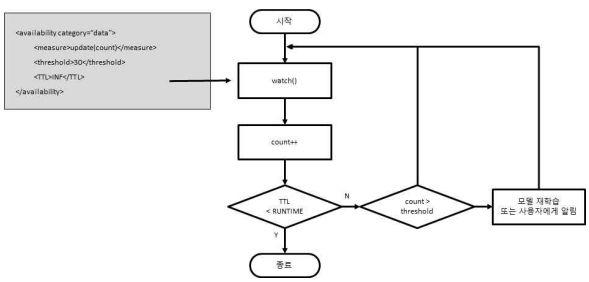

그림 2. 유효성 정의와 유효성에 따른 점진적 학습 과정

표 1은 인텔리전트 데이터베이스 시스템 환경에서 점진적 기계 학습을 수행하기 위하여 적용한 대표적인 점진적 학습 기법의 목록을 보여준다. 본 논문에서 분류를 수행하는 알고리즘에 적용한 Learn++의 경우 데이터를 일정 기간 또는 양으로 구분하여 각각의 작은 분류기를 생성하고 작은 분류기들을 가중치와 함께 병합하여 하나의 큰 분류기를 생성한다. 이러한 점진적 학습 방법은 그래프 스트림 데이터가 입력되는 환경에서 이상치를 감지할 때 사용할 수 있다. 그리고 이때 작은 분류기를 병합할 때 가중치를 부여함에 있어서 모든 작은 분류기의 성능에 따라 다른 가중치를 부여하거나, 최신의 데이터를 가중치가 높게 부여하거나, 중요하다고 생각하는 데이터의 가중치를 높게 부여하는 등 사용자가 다양하게 설정할 수 있다. 또 다른 예로 본 논문에서 의사 결정 트리를 구축하는데 활용한 ID5R는 새롭게 추가되는 데이터를 이용하여 의사 결정 트리를 확장한다. 이전까지의 구축에 사용한 e-score라는 것을 유지하지 때문에 기존 학습대상 데이터를 다시 학습시킬 필요가 없는 방법이다. 사용자는 자신이 서비스에서 활용하고자 하는 모델과 데이터의 특성에 따라 유효성 및 점진적 기법을 설정하여 활용할 수 있다.

표 1. 대표적 점진적 학습 기법

구분	정보
Decision trees	IDE4
	ID5R
Decision rules	Decision rules
Artificial neural networks	RBF networks
	Learn++
	Fuzzy ARTMAP
	TopoART
	IGNG
SVM	incremental SVM

Ⅲ. 결론

본 논문에서는 인텔리전트 데이터베이스 시스템 환경에서 자율 기계 학습을 위한 점진적 학습 기법을 제안하였다. 제안하는 기법은 점진적 학습을 위하여 학습대상 데이터를 이미 학습에 사용된 데이터와 그렇지 않은 데이터로 구분한다. 이와 같이 구분된 학습대상 데이터를 기반으로 새롭게 축적된 데이터만을 활용한 학습 모델과 점진적 학습을 수행한 학습 모델을 도출한다. 마지막으로 제안하는 기법은 점진적 학습을 수행한 모델들을 다양하게 결합하여 성능이 뛰어난 학습 모델을 응용에 활용한다.

향후 연구로는 인텔리전트 데이터베이스 환경에서 딥러닝에 대한 점진적 학습을 지원할 수 있도록 다양한 실제 응용에 활용되는 데이터를 분석하여 지속적으로 점진적 학습 기법을 추가할 예정이다.

ACKNOWLEDGMENT

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-차세대정보ㆍ컴퓨팅기술개발사업의 지원(No. NRF-2017M3C4A7069432), 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. 2019R1A2C2084257), 그리고 과학기술정보통신부 및 정보통신기획평가원의 Grand ICT연구센터지원사업(IITP-2020-1711120023)의 연구결과로 수행되었음

참 고 문 헌

[1] Schlimmer, J. C., and Fisher, D. “A case study of incremental concept induction.”In AAAI. Vol. 86, pp. 496-501, 1986.

[2] Tscherepanow, M., Kortkamp, M., and Kammer, M. “A hierarchical ART network for the stable incremental learning of topological structures and associations from noisy data.” Neural Networks, Vol. 24, No.8, pp. 906-916, 2011.