



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0134308
(43) 공개일자 2019년12월04일

(51) 국제특허분류(Int. Cl.)
G06N 3/04 (2006.01) G06N 3/08 (2006.01)
(52) CPC특허분류
G06N 3/04 (2013.01)
G06N 3/08 (2013.01)
(21) 출원번호 10-2018-0059777
(22) 출원일자 2018년05월25일
심사청구일자 2018년05월25일

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
양지훈
서울특별시 은평구 백련산로 2길 19, 백련산힐스
테이트 1차 105동 603호
임세희
서울특별시 광진구 중곡2동 53-22
(74) 대리인
이지연

전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **합성곱 신경망을 활용한 데이터 확장방법 및 장치**

(57) 요약

본 발명에 따르는 합성곱 신경망을 활용한 데이터 확장방법은, 연속되는 시계열 데이터들을 입력받는 제1단계; 상기 시계열 데이터들 각각에 대해 특징들을 검출하여 시계열 데이터들 각각의 특징들을 2차원 배열로 표현하여 입력 데이터들을 생성하는 제2단계; 상기 입력 데이터들 중 연속된 두개 이상의 입력 데이터를 입력받아 합성곱 신경망 처리하여 새로운 데이터를 생성하는 제3단계; 및 상기 새로운 데이터와 입력 데이터를 결합하여 확장된 데이터를 생성하는 제4단계;를 구비함을 특징으로 한다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호 20150003690031001
 부처명 과학기술정보통신부
 연구관리전문기관 정보통신기술진흥센터
 연구사업명 대학ICT연구센터육성사업
 연구과제명 글로벌 핀테크 벤처 기업 및 창의적 인재 육성을 위한 스마트 핀테크 플랫폼 기술개발
 기 여 율 1/2
 주관기관 서강대학교 산학협력단
 연구기간 2018.01.01 ~ 2018.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 2017-0-00498
 부처명 과학기술정보통신부
 연구관리전문기관 정보통신기술진흥센터
 연구사업명 정보보호핵심원천기술개발
 연구과제명 차분 프라이버시 기반 비식별화 기술 개발
 기 여 율 1/2
 주관기관 서강대학교 산학협력단
 연구기간 2018.01.01 ~ 2018.12.31

명세서

청구범위

청구항 1

합성곱 신경망을 활용한 데이터 확장방법에 있어서,

연속되는 시계열 데이터들을 입력받는 제1단계;

상기 시계열 데이터들 각각에 대해 특징들을 검출하여 시계열 데이터들 각각의 특징들을 2차원 배열로 표현하여 입력 데이터들을 생성하는 제2단계;

상기 입력 데이터들 중 연속된 두개 이상의 입력 데이터를 입력받아 합성곱 신경망 처리하여 새로운 데이터를 생성하는 제3단계; 및

상기 새로운 데이터와 입력 데이터를 결합하여 확장된 데이터를 생성하는 제4단계;를 구비함을 특징으로 하는 합성곱 신경망을 활용한 데이터 확장방법.

청구항 2

제1항에 있어서,

상기 제2단계가,

상기 시계열 데이터들 각각에 대해 특징들을 검출하는 단계;

상기 시계열 데이터들 각각에 대한 특징들의 수가 제공수인지를 판별하는 단계;

상기 시계열 데이터들 각각에 대한 특징들의 수가 제공수가 아니면 제공수만큼의 특징만을 선택하여 특징들의 수를 줄이는 단계;

상기 시계열 데이터들 각각에 대한 제공수의 특징들을 2차원 배열로 변환하여 입력 데이터를 생성하는 단계;로 구성됨을 특징으로 하는 합성곱 신경망을 활용한 데이터 확장방법.

청구항 3

제2항에 있어서,

상기 특징들 중 선택되는 특징들은,

해당 시계열 데이터의 라벨과 상관관계가 높은 특징들임을 특징으로 하는 합성곱 신경망을 활용한 데이터 확장방법.

청구항 4

제1항에 있어서,

상기 제3단계는,

상기 입력 데이터들 중 연속된 두개 이상의 입력 데이터를 입력받아 합성곱 신경망 처리하는 단계;

상기 합성곱 신경망 처리한 결과와 다음 입력 데이터를 비교하여 MSE가 최소화되도록 합성곱 신경망을 학습하는 단계;

학습된 합성곱 신경망에 의해 상기 입력 데이터들 중 연속된 두개 이상의 입력 데이터를 처리하여 새로운 데이터를 생성하는 단계;로 구성됨을 특징으로 하는 합성곱 신경망을 활용한 데이터 확장방법.

청구항 5

합성곱 신경망을 활용한 데이터 확장장치에 있어서,

연속되는 시계열 데이터들을 입력받아, 상기 시계열 데이터들 각각에 대해 특징들을 검출하여 시계열 데이터들

각각의 특징들을 2차원 배열로 표현하여 입력 데이터들을 생성하는 데이터 전처리부;

상기 입력 데이터들 중 연속된 두개 이상의 입력 데이터를 입력받아 합성곱 신경망 처리하여 새로운 데이터를 생성하고, 상기 새로운 데이터와 입력 데이터를 결합하여 확장된 데이터를 생성하는 합성곱 신경망 처리부;를 구비함을 특징으로 하는 합성곱 신경망을 활용한 데이터 확장장치.

청구항 6

제5항에 있어서,

상기 데이터 전처리부가,

상기 시계열 데이터들 각각에 대해 특징들을 검출하고,

상기 시계열 데이터들 각각에 대한 특징들의 수가 제공수인지를 판별하고,

상기 시계열 데이터들 각각에 대한 특징들의 수가 제공수가 아니면 제공수만큼의 특징만을 선택하여 특징들의 수를 줄이고,

상기 시계열 데이터들 각각에 대한 제공수의 특징들을 2차원 배열로 변환하여 입력 데이터를 생성함을 특징으로 하는 합성곱 신경망을 활용한 데이터 확장장치.

청구항 7

제6항에 있어서,

상기 특징들 중 선택되는 특징들은,

해당 시계열 데이터의 라벨과 상관관계가 높은 특징들임을 특징으로 하는 합성곱 신경망을 활용한 데이터 확장장치.

청구항 8

제5항에 있어서,

상기 합성곱 신경망 처리부는,

상기 입력 데이터들 중 연속된 두개 이상의 입력 데이터를 입력받아 합성곱 신경망 처리하고,

상기 합성곱 신경망 처리한 결과와 다음 입력 데이터를 비교하여 MSE가 최소화되도록 합성곱 신경망을 학습하고,

학습된 합성곱 신경망에 의해 상기 입력 데이터들 중 연속된 두개 이상의 입력 데이터를 처리하여 새로운 데이터를 생성함을 특징으로 하는 합성곱 신경망을 활용한 데이터 확장장치.

발명의 설명

기술 분야

[0001] 본 발명은 데이터 확장 기술에 관한 것으로, 더욱 상세하게는 작은 크기이며 변동이 심한 시계열 데이터에 대해서도 예측 성능을 유지할 수 있게 합성곱 신경망을 활용하여 시계열 데이터를 확장하는 합성곱 신경망을 활용한 데이터 확장방법 및 장치에 관한 것이다.

배경 기술

[0002] 시계열 데이터를 예측하는 것은 데이터 분석 및 머신 러닝 분야에서 큰 관심을 받아왔다. 금융, 의학, 생명 등 산업 분야와 무관하게 과거를 통해 미래를 예측하는 것에 관심을 가져왔고, 이는 머신 러닝(Machine Learning), 통계학 등의 학문의 발전을 촉진하였다.

[0003] 최근에는 인공지능이 많은 관심을 받으며 여러 분야에 응용되고 있다. 이러한 알고리즘이 충분히 성능을 발휘하기 위해서는 데이터 집합의 확보가 우선되어야 한다. 그러나 데이터 집합의 크기가 충분히 크지 않으면 과적합

(Overfitting)이 일어날 확률이 높아져 예측 성능을 떨어뜨리기 때문이다.

- [0004] 이에 종래에는 데이터를 확장하여 데이터 집합의 크기를 충분히 크게 만들고 있다. 종래의 데이터 확장 방법을 설명한다. 상기 데이터 확장은 데이터 변환(translating), 데이터 크기 재조정(re-scaling), 데이터 왜곡(distorting) 방식으로 이루어진다. 이러한 방식은 데이터의 라벨(label)에는 영향을 주지 않으면서 새로운 데이터를 생성할 수 있다.
- [0005] 먼저 음성 인식 분야에서의 데이터 확장 과정을 설명한다. 상기 음성 인식 분야에서의 데이터 확장은 짧은 길이의 발화로 신뢰성있는 모델 설계가 어려운 경우에는 성도 길이 변화(vaocal tract length perturbation)을 통해 발화를 확장하였다. 상기 성도는 발성기관의 일부를 일컫는데, 상기 성도의 길이 차이에 따라 발화의 변이가 생긴다. 이러한 변이를 제거하는 방법으로 성도 길이 정규화(vocal tract length normalization) 방법이 있으며, 이를 역으로 적용하여 임의의 변화를 주는 방법을 성도 길이 변화라고 한다. 또한 말 속도를 조절하여 데이터 집합을 확장하는 방법도 존재한다.
- [0006] 그리고 이미지 분야에서의 데이터 확장 과정을 설명하면, 상기 이미지 분야에서 데이터 확장은 자르기, 반전, 색변화 또는 아핀 변환(affine transformation) 등으로 데이터를 확장하였다. 노게이라(nogueira) 등은 이미지의 0.8배 크기만큼 잘라내어 데이터를 확장하는 방법을 사용하였으며, 이는 Nogueira, R. F., de Alencar Lotufo, R., & Machado, R. C. (2014, October). Evaluating software-based fingerprint liveness detection using Convolutional Networks and Local Binary Patterns. In Biometric Measurements and Systems for Security and Medical Applications (BIOMS) Proceedings, 2014 IEEE Workshop on (pp. 22-29). IEEE.에 개시되었다. 그리고 krizhevsky 등은 이미지 반전과 RGB의 색변화를 통하여 데이터를 확장하였으며, 이는 Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems(pp. 1097-1105).에 개시되었다. 그리고 ciregan은 이미지를 무작위 이동(random translation), 회전 등을 적용하여 데이터를 확장하였으며, 이는 Ciregan, D., Meier, U., & Schmidhuber, J.(2012, June). Multi-column deep neural networks for image classification. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp.3642-3649). IEEE.에 개시되었다. 또한 아핀변환(affine transformation)과 왜곡을 활용하여 데이터를 확장하는 방법도 개시되어 있으며, 이는 Cire, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J.(2011). High-performance neural networks for visual object classification. arXiv preprint arXiv:1102.0183. Simard, P. Y., Steinkraus, D., & Platt, J. C.(2003, August). Best practices for convolutional neural networks applied to visual document analysis. In ICDAR (Vol. 3, pp. 958-962).에 개시되었다.
- [0007] 상기한 바와 같은 노력에도 불구하고 시계열 데이터에 대한 데이터 확장은 용이하지 않았다. 이는 시계열 데이터의 특성상 새로운 지표가 등장하거나 신규 시계열 데이터가 등장하는 경우가 잦기 때문이다.
- [0008] 즉 주식 데이터나 전력 예측 데이터 등과 같은 시계열 데이터는 회계 지표가 새로이 등장하거나 새로운 가전 기기가 포함되는 등 다양한 경우가 존재하는 작은 크기의 변동이 심한 시계열 데이터의 경우에는 데이터 예측시에 데이터 모델에 과적합되어 데이터 예측 성능이 저하될 수 있는 원인이 되었다.
- [0009] 이에 종래에는 작은 크기이며 변동이 심한 시계열 데이터에 대해서도 예측 성능을 유지할 수 있게 데이터 모델에 대한 데이터 확장을 이행할 수 있는 기술의 개발이 절실하게 요망되었다.
- [0010]

선행기술문헌

특허문헌

- [0011] (특허문헌 0001) 한국특허공개 제1020160102690호
(특허문헌 0002) 한국특허공개 제1020180028893호
(특허문헌 0003) 한국특허공개 제1020170134508호

발명의 내용

해결하려는 과제

- [0012] 본 발명은 작은 크기이며 변동이 심한 시계열 데이터에 대해서도 예측 성능을 유지할 수 있게 합성곱 신경망을 활용하여 시계열 데이터를 확장하는 합성곱 신경망을 활용한 데이터 확장방법 및 장치를 제공하는 것을 목적으로 한다.

과제의 해결 수단

- [0013] 상기한 목적을 달성하기 위한 본 발명에 따르는 합성곱 신경망을 활용한 데이터 확장방법은, 연속되는 시계열 데이터들을 입력받는 제1단계; 상기 시계열 데이터들 각각에 대해 특징들을 검출하여 시계열 데이터들 각각의 특징들을 2차원 배열로 표현하여 입력 데이터들을 생성하는 제2단계; 상기 입력 데이터들 중 연속된 두개 이상의 입력 데이터를 입력받아 합성곱 신경망 처리하여 새로운 데이터를 생성하는 제3단계; 및 상기 새로운 데이터와 입력 데이터를 결합하여 확장된 데이터를 생성하는 제4단계;를 구비함을 특징으로 한다.

발명의 효과

- [0014] 본 발명은 합성곱 신경망을 활용하여 시계열 데이터를 확장하며 작은 크기이며 변동이 심한 시계열 데이터에 대해서도 예측 성능을 유지할 수 있게 하는 효과를 야기한다.

도면의 간단한 설명

- [0015] 도 1은 본 발명에 따르는 합성곱 신경망을 활용한 데이터 확장장치의 구성도.
 도 2는 도 1의 데이터 확장부의 상세구성도.
 도 3은 도 2의 데이터 전처리부의 처리과정도.
 도 4는 도 2의 합성곱 신경망 처리부의 처리과정도.
 도 5는 도 4의 합성곱 신경망 처리부에 의한 데이터 처리를 예시한 도면.
 도 6은 도 1의 데이터 예측부의 구성도.
 도 7 내지 도 9는 본 발명에 따르는 성능 실험 결과를 도시한 도면.

발명을 실시하기 위한 구체적인 내용

- [0016] 본 발명은 합성곱 신경망을 활용하여 시계열 데이터를 확장하며 작은 크기이며 변동이 심한 시계열 데이터에 대해서도 예측 성능을 유지할 수 있게 한다.
- [0017] 상기한 본 발명의 바람직한 실시예에 따르는 합성곱 신경망을 활용한 학습 데이터 확장장치를 도면을 참조하여 상세히 설명한다.
- [0018] <합성곱 신경망을 활용한 학습 데이터 확장장치의 구성>
- [0019] 도 1은 본 발명의 바람직한 실시예에 따르는 합성곱 신경망을 활용한 학습 데이터 확장장치의 구성을 도시한 것이다.
- [0020] 상기 합성곱 신경망을 활용한 학습 데이터 확장장치는 데이터 확장부(100)와 데이터 예측부(200)로 구성된다.
- [0021] 상기 데이터 확장부(100)는 연속된 시점의 학습 데이터인 시계열 데이터를 전처리하여 합성곱 신경망에 입력하고, 상기 합성곱 신경망은 상기 전처리된 시계열 데이터를 입력받아 확장하여 출력한다.
- [0022] 상기 데이터 예측부(200)는 상기 학습 데이터 확장부(100)에 의해 확장된 시계열 데이터를 이용하여 집합을 생성하고, 여러개의 LSTM을 사용하여 앙상블하여 예측결과를 생성하여 출력한다.
- [0023] <합성곱 신경망을 활용한 학습 데이터 확장부(100)의 구성 및 동작>
- [0024] 상기 학습 데이터 확장부(100)는 데이터 전처리부(102)와 합성곱 신경망 처리부(104)와 MSE 최소화 처리부(106)로 구성된다.
- [0025] 상기 데이터 전처리부(102)는 훈련 데이터인 시계열 데이터를 제공받아 합성곱 신경망 처리부(104)에서 처리 가

능한 2차원 배열 형태로 변환한다.

[0026] 상기 데이터 전처리부(102)의 처리과정을 도시한 것이 도 3이다. 상기 도 3을 참조하여 상기 데이터 전처리부(102)의 동작을 설명한다.

[0027] 상기 데이터 전처리부(102)는 시계열 데이터를 입력받아 특징들로 표현한다(300단계).

[0028] 좀더 설명하면, 상기 시계열 데이터는 시간순으로 색인화된 일련의 데이터 요소이며, 상기 시계열 데이터가 수집된 기간이 T라면 시계열 데이터 집합은 수학식 1과 같이 나타낼 수 있다.

수학식 1

[0029]
$$\mathbf{x} = [x_t | t = 1, \dots, T]$$

[0030] 상기 수학식 1에서 $\mathbf{x}$ 는 시계열 데이터를 나타내며, x_t 는 t 시각에서의 데이터를 나타낸다.

[0031] 그리고 상기 데이터의 특징을 f 라 하고, 데이터가 N 개의 특징으로 구성되어 있다면 상기 데이터는 수학식 2로 나타낼 수 있다.

수학식 2

[0032]
$$x = [f_n | n = 1, \dots, N]$$

[0033] 상기 데이터 전처리부(102)는 상기 데이터의 특징 N 이 제공수인지를 판별하고(302단계), 상기 데이터의 특징 N 이 제공수가 아니면 이 데이터를 2차원 구조 배열로 바꾸기 위해 특징 선택(feature selection)을 통해 N 보다 작은 제공수로 변환한다.

[0034] 상기 데이터 전처리부(102)는 입력 데이터의 특징들 $f_i (i = 1, 2, \dots, N)$ 각각과 라벨의 피어슨 상관계수(pearson correlation coefficient)를 계산한다(304단계).

[0035] 상기 피어슨 상관계수는 수학식 3에 따라 계산된다.

수학식 3

[0036]
$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}.$$

[0037] 여기서, 입력 데이터가 주식 데이터라 가정하면, 상기 입력 데이터는 거래량, 가격변동 등으로 구성되는 특징 f_i 와 주식의 등락(+,-) 등의 라벨로 구성된다. 상기 수학식 3에서 x_i 는 $f_i (i = 1, 2, \dots, N)$ 이며, y_i 는 라벨값이다.

[0038] 상기 데이터 전처리부(102)는 입력 데이터의 특징들 $f_i (i = 1, 2, \dots, N)$ 각각과 라벨의 피어슨 상관계수(pearson correlation coefficient)의 계산결과를 토대로, 입력 데이터들의 특징들 중 미리 정해둔 수, 즉 제공수만큼 상관관계가 높은 특징들을 선택한다(306단계). 여기서 상기 특징들의 선택시에는 피어슨 상관계수 계산결과 중 양의 상관관계나 음의 상관관계에 무관하게 상관관계가 높은 상위 특징들이 선택된다. 그리고 상기 제공수는 N 보다 작은 제공수 중 가장 큰 수인 K 개로 선택된다. 이에 따라 데이터 x 는 $x = [f_n | 1, \dots, K]$ 로 표현된다. 그리고 임의의 자연수 k 에 대해 $K = k^2$ 이므로 데이터 x 는 가로, 세로 크기가 k 인 2차원 배열로 표현될 수 있으므로, 상기 데이터 전처리부(102)는 입력 데이터의 특징들을 2차원 배열로 표현하여 합성곱 신경망에 적합한 신경망에

적합한 입력 데이터로 변환한다(308단계).

- [0039] 상기한 바와 달리 상기 특징들의 수가 제곱수이면, 데이터 전처리부(102)는 입력 데이터의 특징들을 2차원 배열로 표현하여 합성곱 신경망에 적합한 신경망에 적합한 입력 데이터로 변환한다(308단계).
- [0040] 상기 데이터 전처리부(102)는 상기한 데이터 전처리 과정을 모든 데이터 $x_t(t=1, \dots, T)$ 에 적용하여 합성곱 신경망에 적합한 입력 데이터 집합 X^{img} 를 생성한다.
- [0041] 상기 합성곱 신경망 처리부(104) 및 MES 최소화 처리부(106)는 상기 입력 데이터 집합 X^{img} 를 입력받아 확장하여 출력한다.
- [0042] 먼저 합성곱 신경망을 이용하여 데이터를 확장하는 전체 과정을 간략하게 설명한다.
- [0043] 이미지의 분류를 목표로 하는 합성곱 신경망과 다르게 본 발명의 합성곱 신경망은 시계열 데이터의 확장을 목표로 하므로, 연속된 k개의 데이터를 입력으로 하여 생성한 합성곱 신경망의 결과는 입력 데이터의 다음 데이터인 (k+1)번째 데이터와 유사해야 한다. 이에 본 발명은 비용함수에 반영하여 합성곱 신경망이 시계열 데이터의 정보를 파악할 수 있도록 하여 시계열 데이터를 훈련한다.
- [0044] 상기 시계열 데이터의 훈련이 끝난 후, 기존 데이터 집합을 다시 합성곱 신경망에 입력으로 집어넣어 기존 데이터 집합과 동일한 기간의 데이터 집합을 새로 생성할 수 있다.
- [0045] 이러한 과정을 거쳐 원본 데이터 집합과 유사한 새로운 데이터 집합의 생성하여 데이터 확장을 이룰 수 있다.
- [0046] 이러한 합성곱 신경망 처리부(104) 및 MES 최소화 처리부(106)의 처리과정을 도 4를 참조하여 상세히 설명한다.
- [0047] 이미지 처리를 위한 합성곱 신경망은 이미지의 크기에 따른 폭(width)과 높이(height)와 컬러에 따른 깊이(depth)가 입력되므로, 상기 합성곱 신경망 처리부(104)는 입력 데이터 집합 X^{img} 의 크기에 따라 폭(width)과 높이(height)를 결정하고, 연속된 시점의 데이터 쌍의 개수를 깊이(depth)로 결정한다(400단계). 여기서, 상기 입력 데이터 집합 X^{img} 이 $k \times k$ 와 같은 2차원 배열이라면 입력 데이터는 $k \times k \times 2$ 구조이며, 이는 도 5의 (a)에 도시한 바와 같다.
- [0048] 상기한 바와 같은 과정을 거쳐 합성곱 신경망의 입력 데이터 집합 X^{tr} 를 구성하면 수학적 식 4와 같이 표현할 수 있다.

수학적 식 4

[0049]
$$X^{tr} = [(x_t^{img}, x_{t+1}^{img}) | t = 1, \dots, T-1]$$

- [0050] 상기 수학적 식 4에서 X^{tr} 은 합성곱 신경망의 입력 데이터 집합이고, 상기 순차적인 입력 데이터 $(x_t^{img}, x_{t+1}^{img})$ 는 순차적인 입력 데이터 쌍을 나타내며, T는 시계열 데이터의 전체 시간을 나타낸다.
- [0051] 상기 합성곱 신경망 처리부(104)는 상기 입력 데이터 쌍인 $(x_t^{img}, x_{t+1}^{img})$ 를 입력으로 받아 합성곱 신경망 처리하여 MSE 최소화 처리부(106)로 출력한다(402단계).
- [0052] 상기 MSE 최소화 처리부(106)는 상기 합성곱 신경망의 결과와 다음 시점, 즉 (t+2) 시점의 데이터 x_{t+2}^{img} 와 비교하여 MSE가 최소화되도록 합성곱 신경망을 훈련한다(404단계).
- [0053] 좀더 설명하면, 상기 (t+2) 시점의 데이터 집합 $X^{compare}$ 은 수학적 식 5와 같이 구성될 수 있다.

수학식 5

$$X^{compare} = [x_{t+2}^{img} | t = 3, \dots, T]$$

[0054]

[0055] 즉, X^{tr} 는 입력 데이터이고, $X^{compare}$ 는 라벨값으로 합성곱 신경망에 적용된다.

[0056] 상기 데이터 확장을 위한 합성곱 신경망 구조는 일반적인 합성곱 신경망과 유사하지만 마지막 결과(output)를 출력하는 계층이 다르다. 일반적인 이미지를 분류하는 합성곱 신경망의 경우 마지막 계층에서 소프트맥스(softmax)를 사용하여 가장 큰 값을 가진 클래스(class)로 분류한다. 그러나 본 발명에서는 새로운 데이터를 생성하는 것이 목적이므로, 소프트맥스 계층을 사용하지 않고 입력 데이터와 같은 폭(width)과 높이(height)를 갖는 $[width \times height \times 1]$ 구조를 결과 값으로 받을 수 있도록 합성곱 신경망의 구조가 결정된다. 예를 들어, 입력 데이터가 $k \times k \times 2$ 구조라면, 결과 값은 $k \times k \times 1$ 구조가 되도록 설계한다.

[0057] 그리고 합성곱 신경망의 결과와 $X^{compare}$ 의 데이터의 차이가 적도록 비용함수를 설정하여 합성곱 신경망을 훈련시킨다. 여기서 합성곱 신경망의 결과와 $X^{compare}$ 의 차이는 MSE(Mean Squared Error)로 계산한다. 상기 MSE는 수학식 6에 따라 계산되며, 이를 비용함수로 설정한다.

수학식 6

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

[0058]

[0059] 상기 수학식 6에서 Y_i 는 합성곱 신경망의 결과값이고 $\hat{Y}_i$ 는 $X^{compare}$ 이다.

[0060] 상기의 합성곱 신경망의 훈련이 끝나면, 상기 합성곱 신경망 처리부(104)는 기존 데이터 집합을 훈련된 합성곱 신경망 처리하여 데이터를 생성한다(404단계). 이후 상기 합성곱 신경망 처리부(104)는 기존 데이터 집합에 새로운 데이터를 추가하여 데이터를 확장한다(406단계).

[0061] 이러한 데이터 확장 과정은 도 5의 (b)에 도시한 바와 같다. 상기 도 5의 (b)를 참조하면, 상기 합성곱 신경망 처리부(104)는 입력 데이터 $x_1^{img}, x_2^{img}, \dots, x_{n-1}^{img}$ 를 입력받아 $x'_3, x'_4, \dots, x'_n$ 가 생성되어 출력한다. 이로서 입력 데이터는 $x_1^{img}, x_2^{img}, \dots, x_{n-1}^{img}, x'_3, x'_4, \dots, x'_n$ 로 확장된다. 이렇게 확장된 데이터들은 $k \times k$ 의 2차원 배열 형태이다.

[0062] 이후 상기 합성곱 신경망 처리부(104)는 $k \times k$ 의 2차원 배열 형태의 확장된 데이터들을 실제 예측을 위해 $1 \times k^2$ 의 형태로 변환하여 최종 확장된 데이터 집합을 구성하여 출력한다(408단계).

[0063] 상기한 본 발명의 바람직한 실시예에서는 연속된 두 데이터를 이용하여 새로운 데이터를 생성하였으나, 연속된 두개 이상의 데이터를 이용하여 새로운 데이터를 생성할 수도 있으며, 이는 본 발명에 의해 당업자에게 자명하다.

[0064] 상기한 데이터 확장부(100)가 생성한 확장된 데이터 집합은 시계열 데이터를 예측하는 데이터 예측부(200)로 제공된다.

[0065] <시계열 데이터를 예측하는 데이터 예측부(200)의 구성 및 동작>

[0066] 도 6은 시계열 데이터를 예측하는 데이터 예측부(200)의 구성을 도시한 것이다. 상기 도 6을 참조하면, 상기한 시계열 데이터를 예측하는 데이터 예측부(200)는 제1 내지 제N LSTM(Long Short-Term Memory Networks) 예측부

(5001~500N)와 앙상블부(502)로 구성된다.

- [0067] 상기 제1 내지 제N LSTM부(5001~500N)는 원본 데이터 집합과 새로이 생성된 확장된 데이터 집합들의 개수에 대응되게 구비되며, 상기 제1 내지 제N LSTM부(5001~500N) 각각은 입력된 데이터 집합에 대해 예측을 이행하고 그 결과를 출력한다. 상기 제1 내지 제N LSTM부(5001~500N)는 RNN(Recurrent Neural Network)의 히든 노드에 셀 상태(cell-state)를 추가한 구조를 가진다.
- [0068] 상기 제1 내지 제N LSTM부(5001~500N) 각각의 출력은 앙상블 처리부(504)에 입력된다. 상기 앙상블 처리부(504)는 상기 제1 내지 제N LSTM부(5001~500N) 각각의 출력의 평균을 취하여 최종 예측 결과를 출력한다.
- [0069] 상기 앙상블 처리부(504)는 상기 제1 내지 제N LSTM부(5001~500N)에서 나온 값으로부터 다수결 판단을 적용하여 최종 예측 결과를 출력한다.
- [0070] <실험 및 결과>
- [0071] 상기한 바와 같이 구성되는 본 발명의 바람직한 실시예에 따르는 합성곱 신경망을 활용한 데이터 확장장치의 성능을 증명하기 위해 실험을 이행하였으며 그 결과는 다음과 같다.
- [0072] 먼저 본 발명의 성능을 증명하기 위한 데이터로는 주식 수익률 데이터, 이더리움 가격 데이터 및 기기의 에너지 사용량 데이터를 사용하였다.
- [0073] 상기 주식 수익률 데이터는 2000년 1월 ~2016년 12월까지 월별 주식의 수익률과 195개의 특징으로 구성되며, 월별로 200개의 주식 정보가 수집되었고 매월마다 구성되는 주식 종목에는 변동이 있다. 또한 상기 195개의 특징들도 전체 기간동안 수집되지 않아 실제 활용할 수 있는 데이터는 기간이 매우 짧다. 따라서 본 실험에서는 주식별로 데이터를 따로 분리하여 실험을 진행하였다.
- [0074] 그리고 상기 이더리움 가격 데이터는 kaggle을 통해 얻을 수 있으며, 총 특징의 개수는 15개이고 789행으로 이루어져 있고, 라벨은 이더리움의 가격이다.
- [0075] 그리고 기기의 에너지 사용량 데이터는 가전 기기의 에너지 데이터는 Candanedo, L. M., Feldheim, V., & Deramaix, D. (2017). Data driven prediction models of energy use of appliances in a low-energy house. Energy and Buildings, 140, 81-97.을 통해 얻을 수 있었다. 상기 기기의 에너지 사용량 데이터의 특징은 27개이고 75,030행으로 이루어져 있고, 라벨은 에너지 사용량이다.
- [0076] 이제 실험방법에 대해 설명한다. 본 발명에 따른 방법과 랜덤 노이즈 확장 기법을 이용하여 데이터를 확장하는 실험을 이행하였다. 상기 랜덤 노이즈 확장 기법은 일반적으로 적용할 수 있는 임의의 난수를 더해주는 방식이 있다. 이 방식은 원본 데이터에 노이즈를 더하는 방식으로 여러 분포를 통해 노이즈를 생성할 수 있다. 특히 본 실험에서는 각 특징의 표준편차에 해당하는 가우시안 분포를 사용하여 노이즈를 생성하였다. 또한 생성된 노이즈는 $\gamma(0 < \gamma < 1)$ 을 곱하여 조정한 후 더해주었다.
- [0077] 상기한 실험방법에 따른 실험결과를 설명한다. 본 발명에 따른 방법과 랜덤 노이즈 확장 방법을 실험한 결과는 도 7에 도시한 표 1과 같다. 상기 표 1을 참조하면 원본 데이터 집합만 사용하였을 때보다 합성곱 신경망을 통해 데이터를 확장하고 그 데이터 집합을 포함한 실험이 더 좋은 결과를 보인 것을 나타낸다. 즉, 본 발명에 따라 데이터 확장 방법을 이용했을 때와 랜덤 노이즈(Random Noise)를 이용해 데이터 확장을 했을 때 학습한 결과와 실제 라벨(label)사이의 오차 값을 나타낸다. 즉, 오차가 작을수록 더 좋은 결과를 나타냈다고 볼 수 있다. 그리고 Metric이란 오차를 계산할 때 사용한 비용함수를 말하는 것이며, Stock return data와 Ethereum price data에서는 MSE(Mean squared Error)를 사용하였고, Energy use data에서는 RMSE(Root Mean Square Error)를 사용하였다. 그리고 RMSE는 MSE에 root를 계산한 값이다.
- [0078] 그리고 데이터 집합의 개수에 따른 성능은 도 7에 도시한 표 2와 같다. 확장 데이터 1은 입력으로 들어가는 연속된 시점의 데이터를 2개 선택했을 때, 확장 데이터 2는 연속된 시점의 데이터를 3개 선택했을 때, 확장 데이터 3과 4는 연속된 시점의 데이터를 4, 5개 선택한 결과로 생성된 확장 데이터 집합이다. 즉, 원본데이터만 사용하였을 때와, 순차적으로 확장데이터 1, 2, 3, 4까지 더한 결과로서, 표 2와 같이 주식 수익률과 이더리움 가격 데이터는 MSE를 에너지 사용량의 경우 RMSE를 사용한 결과 값이다.
- [0079] 그리고 에너지 데이터의 경우 일반적인 시계열 예측 알고리즘을 추가하여 실험하였고, 타 알고리즘 대비 성능이 좋은 것을 확인하였다.

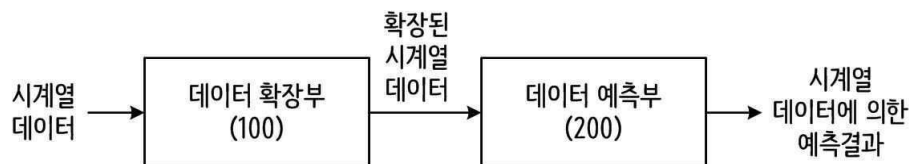
- [0080] 확장 데이터 집합을 포함했을 경우와 포함하지 않을 경우와 더불어 4개의 알고리즘을 실험한 결과는 도 8에 도시한 표 3과 같고, 확장 데이터를 포함한 LSTM이 가장 좋은 성능을 보인다.
- [0081] 이러한 실험을 통해 임의의 노이즈를 더한 데이터 확장 기법과 합성곱 신경망을 통한 데이터 확장 기법을 비교하였고, 합성곱 신경망을 통해 데이터를 확장한 경우가 성능이 더 좋은 것을 확인할 수 있었다. 이를 통해, 데이터를 확장할 때 데이터의 정보를 담은 확장이 필요함을 알 수 있다.
- [0082] 그리고 각각의 확장된 데이터 집합과 원본 데이터 집합의 성능을 비교했다. 이를 통해 입력의 깊이(DEPTH)가 성능에 영향을 미치는 것을 확인할 수 있었다. 입력으로 들어가는 연속된 데이터 개수가 2개 이하일 때는 원본 데이터보다 보통 좋은 성능을 보였다.
- [0083] 또한 원본 데이터와 양상불하는 확장된 데이터 집합의 개수를 달리하며 예측하였을 때에도, 원본 데이터와 확장 데이터 집합 1,2개를 사용하였을 때 좋은 성능을 보였다.
- [0084] 이는 확장된 데이터 집합의 MSE와 연관지어 각 모델의 성능과 유사한 성능을 보인 것으로 해석된다.
- [0085] 이러한 양상불을 사용한 결과를 통해 확장된 데이터 집합이 원본 데이터보다 시계열 정보를 더 포함하고 있음을 확인할 수 있다.
- [0086] 이러한 본 발명은 주식 데이터와 같이 특징들을 수집할 수 있는 유효한 기간이 짧아 데이터 집합이 작아질 때, 이를 확장하여 더 좋은 성능을 낼 수 있다. 이러한 측면은 주로 개장, 폐장 가격, 거래 볼륨(Volume) 등을 특징으로 한 기존 모델들과는 다르게 더 많은 경제지표를 특징으로 활용할 수 있고, 더 좋은 성능을 보일 수 있는 가능성을 열어준다.

부호의 설명

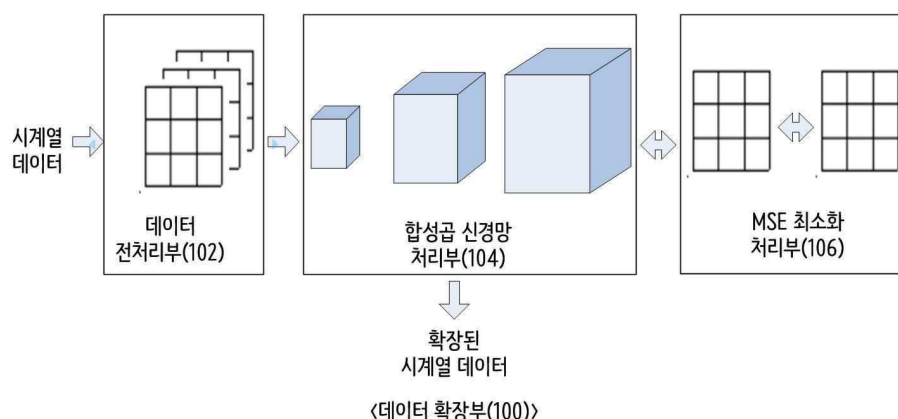
- [0087] 100 : 데이터 확장부
200 : 데이터 예측부

도면

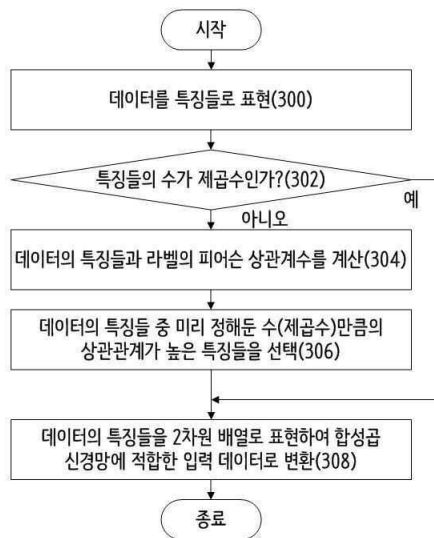
도면1



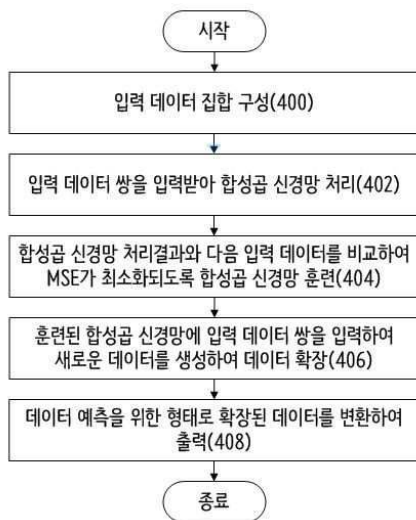
도면2



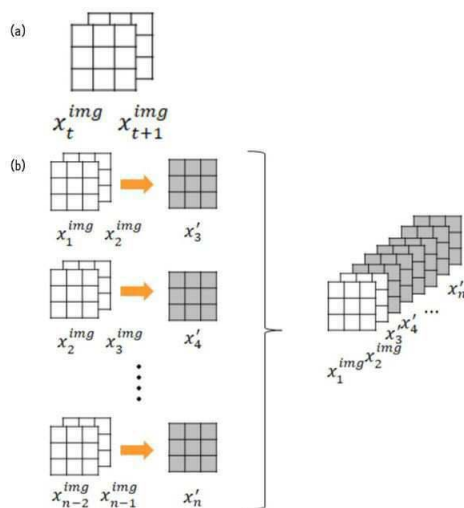
도면3



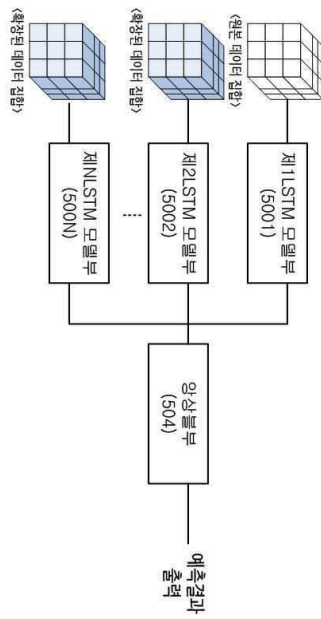
도면4



도면5



도면6



도면7

〈표 1〉 Experimental results compared with only using original data set

Data set	Augmentati on Method	Original data set	Original + Augmented Data set	Metric
Stock return data	CNN	0.040	0.036	MSE
	Random Noise		0.041	
Ethereum price data	CNN	0.015	0.013	RMSE
	Random Noise		0.014	
Energy use data	CNN	67.44	63.83	RMSE
	Random Noise		67.42	

도면8

〈표 2〉 . Performance by number of data sets

Data set	The number of dataset				
	1 (Original data)	2	3	4	5
Stock return data	0.04	0.037	0.036	0.038	0.038
Ethereum price data	0.015	0.013	0.013	0.015	0.016
Energy use data	67.44	66.12	63.83	66.92	68.00

도면9

표 3) RMSE of various algorithm contained proposal augmented algorkthm

Algorithm	RMSE
Linear Regression	93.18
Support Vector Machine	70.74
Random Forest	68.48
Gradient Boosting Machine	66.65
LSTM	67.44
LSTM with data augmentation	63.83