Median Housing value prediction using Conv1D neural network

By:Preetkumar Patel
Lakehead University

I. ABSTRACT

Convolutional neural network is a type of neural networks which are used in image recognition and classification of data. In Conv1D, the kernel slides along one dimension. Basically, time series data or sensory data are used for this method as kernel can move in one dimension along axis of time. This paper details about predicting median housing value using longitude, latitude, housing median age, total number of rooms, total number of bedrooms, population, number of households, and median income attributes of the California housing dataset. The method used for the prediction of house median value is Conv1D neural networks. Finally, we are evaluating the model on testing dataset to check how it performance. Moreover, this paper also discuss about the previously implemented research work which uses the same approach on different dataset.

Keywords: Conv1D, ReLU, medianHouseValue, max pooling, feed forward, LSTM, MSE, R2 Score

II. INTRODUCTION

Nowadays, Real estate is one of the major concerns of the people. Due to the demands for accommodation, people's attention to the price of housing and its median value continues to increase. Providing accurate predictions of the house prices and median value of the house is important task.

Basically CNNs are just multiple layers of convolutions which consist of nonlinear activation functions such as ReLU or tanh applied to results. In a typical neural feedforward network we connect each input neuron to the next layer of each output neuron. However, cnn uses the convolutions over the input layer to calculate the output. Here, each input region is connected to the output neuron. Each layer will apply the filters and finally combines the results. CNN consist of three main layers: convolutional, pooling layer and fully connected layers. The three important layers of CNN are illustrated in the Fig 1. Here, the convolutional layers are stacked, followed by layer pooling in a repeated manner before forwarding to fully-connected layers. The convolutional layer is used to calculate the output of the neurons which are connected to the local input regions by measuring the scalar product between their weights and the region connected to the input volume. The aim of the rectified linear unit (commonly known as ReLu) is to apply the 'element-wise' activation function, such as sigmoid, to the output of the activation generated by the previous layer[2]. The pooling layer then simply downsample along the spatial dimensionality of the input provided, further it helps to reduce the number of parameters within that activation [2]. The fully-connected layers perform the same tasks as those found in the regular ANNs and it will attempt to produce class scores from the activations to be used for classification. It is also suggested that ReLu can be used between these layers to improve performance. [2].

The proposed approach which is Conv1D neural network gives good result on the California housing dataset.In this dataset, the house median value is predicted using other attributes of the dataset.The performance of the model is evaluated using MSE and R2 scores.

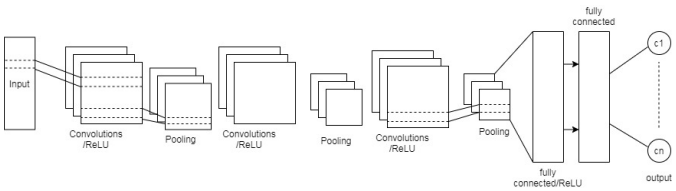

Fig. 1. CNN architecture with convolutional layers which are continuously stacked between ReLus before passing through the pooling layer, before going between one or more fully connected ReLus.

III. RELATED WORK

Many researches have perform methodology of convolutional neural network on different datasets which also shows good results. One of the paper is "Stock Price Prediction on Daily Stock Data using Deep Neural Networks"[3]. In this paper, different deep neural network techniques are used to predict the price of stocks. The methods includes convolutional neural network, long short term memory network and conv1D LSTM are used in order to predict prices of Tata Consultancy Services (TCS) and Madras Rubber Factory Limited (MRF) stocks on a short term basis. In this paper, a deep neural network Conv1D-LSTM is proposed which is based on the combining of layers of two different techniques: CNN and LSTM top redict the price of a stock. The performance of the models is evaluated using RMSE, MAE and MAPE. These errors in Conv1D-LSTM model are found to be very low compared to CNN LSTM.Predictions are mainly categorized as long term and short term. In this paper, it deals with a specific time-series prediction related to the financial sector called Stock Price

Prediction. The variable in this Time-Series is the stock price. Economic benefits can easily be achieved by forecasting the growth of financial structures such as Stocks. Stock activity is highly volatile and very complex in nature[3]. The techniques involved in the time series analysis of financial and stock data have become increasingly important due to their purpose of helping to increase profits while trying to keep a low risk potential [3].All in all, the methodology proposed in this paper is based on the combination of the layers of different techniques into a single deep neural network, while using fewer training features.

Furthermore, the other paper "Forecasting Stock Prices from the Limit Order Book using Convolutional Neural Networks" also gives good results[4]. In this paper, a deep learning methodology, based on Convolutional Neural Networks (CNNs), that predicts the price movements of stocks, using as input large-scale, high-frequency time-series derived from the order book of financial exchanges[4]. The dataset used here contains more than 4 million limit order events and it is also compared with other methods, like Multilayer Neural Networks and Support Vector Machines, shows that CNNs are better suited for this kind of task[4].

The main application of CNN is image classification.The basic approach for the image classification proposed in "Simple convolutional neural network on image classification".The approach uses minister and cifar-10 benchmarking datasets. The author explored different learning rate set methods and different optimization algorithms to solve the optimal parameters of the effect on image classification, based on the Convolutional neural network.It also verifies that the shallow network has a relatively good recognition effect.

IV. Dataset Discription

In the proposed approach California Housing dataset[1] is used in which house median value is predicted using other nine given attributes. This dataset contains 20,640 observations and 9 variables among which 8 are continuous and one of them is categorical attribute. The attributes of the given dataset are:

1) houseMedianAge: This attribute indicates median age of the house in a block, higher the value, older the house.
2) totalRoom: This is total number of rooms within a block.
3) totalBedrooms: Total number of bed rooms within a block.
4) population: Total number of people living in the area.
5) households: Total number of households, a group of residents of a household unit, a block.
6) medianIncome: Median household income within household blocks (US dollars).
7) medianHouseValue: Median house value (measured in US dollars) for households within a block.
8) oceanProximity: Location of the house with respect to sea.
9) longitude: It measures how far the house is from the west.
10) latitude: It measures how far the house is from the north.

This dataset was published in a 1997 in the paper Sparse Spatial Auto regressions by Pace, R. Kelley and Ronald Barry, which was published in the Journal of Statistics and Probability Letters. It was developed using data from the 1990 California census. This dataset contains one row per census block group. A block group usually has a population of 600-3,000[1].

V. Proposed Methodology

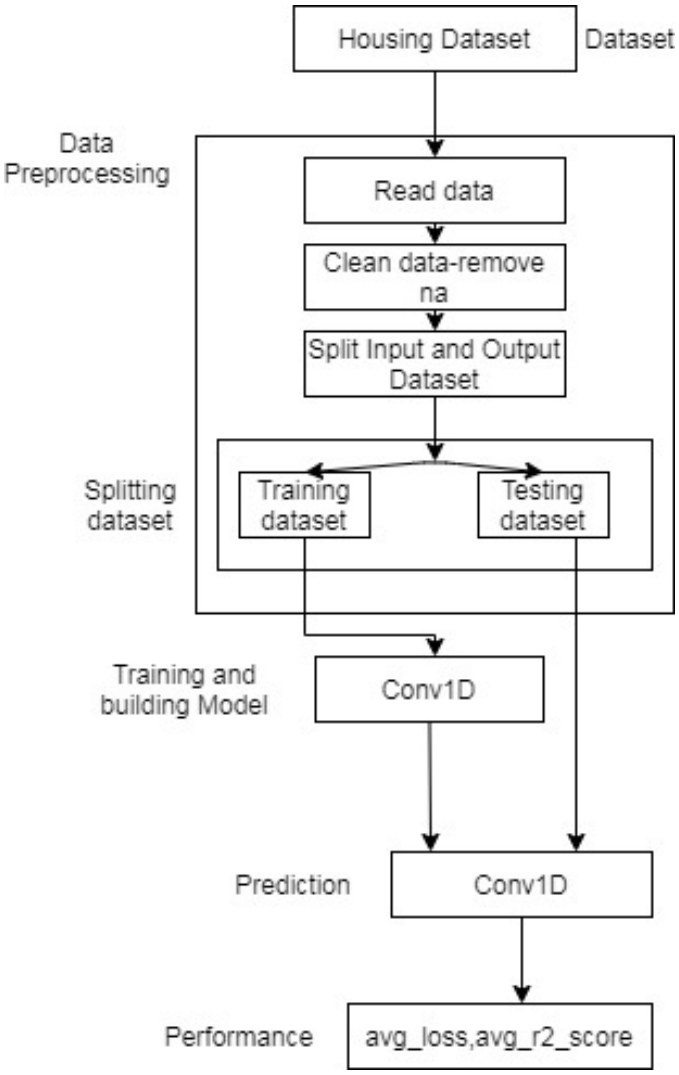

Fig. 2. Flow diagram of proposed method

The proposed methodology uses California housing dataset for prediction of housing median value using other attributes. To design the 1D CNN model for predicting the housing median value on the input dataset some basic steps are performed such as data collection, data preprocessing, building and training the model, and evaluating the model as shown in the Fig 2. First read the dataset using readcsv function. After data collection, next step is to do preprocessing step. The preprocessing step includes data cleaning that is

removal of all NAs in the dataset. After removal of the incomplete entities, each feature of the dataset is plot on separate sub-plots as demonstrated in the Fig 3. This is done using dropna function. Then the dataset is spilt into input and output data. Input data is from column longitude to medianIncome column and the output data is medianHouseValue column. Moreover, the dataset is split into 80% for training the dataset and 20% for testing purpose.

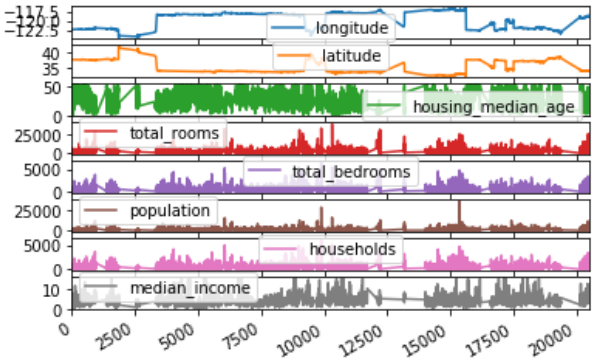

Fig. 3. Subplots of dataset in a single figure

After preprocessing step, the conv1D model is defined and trained on the training dataset. This is used to predict the house median value of the given dataset. The CnnRegressor function used in the approach defines input layer, max pooling layer, convolutional layer, flatten layer, linear layer, output layer. Furthermore, feed function is defined to take inputs from the model. In this function, output of the first layer is obtained and run through the ReLU activation function. After this,the output of the second layer is obtained and again run it through the ReLU activation function. Now, get the output of the flatten layer and linear layer and run it with activation function. Moreover, obtain the output of the output layer and return the output.Now, model loss function is defined which returns the L1 loss and R2 Score of he passed model on passed dataloader. Finally, model is trained and in this way, the steps to train the model are completed.

Furthermore, after completion of the training, the models are used for obtaining the predictions on the test set. The model is then evaluated on the basis of the outcome from the prediction. The networks are then evaluated using MSE as loss function and R2 score. After evaluation of the model, the scores are 62634.1 and 0.43 respectively.

VI. Conclusion

The paper discuss about predicting median housing value using the longitude, latitude, median housing age, total number of rooms, total number of bedrooms, population, number of households, and median income attributes of the California Housing Dataset. The methodology used to predict house median value is the Conv1D neural network. Moreover, the performance of the model is evaluated using MSE (mean square error) that is L1 loss and R2 scores, which are 62634.1 and 0.43 respectively.

References

[1] https://github.com/ageron/handson-ml/tree/master/datasets/housing

[2] O'Shea, Keiron & Nash, Ryan. (2015). An Introduction to Convolutional Neural Networks. ArXiv e-prints.

[3] S. Jain, R. Gupta and A. A. Moghe, "Stock Price Prediction on Daily Stock Data using Deep Neural Networks," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), Bhopal, India, 2018, pp. 1-13.

[4] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj and A. Iosifidis, "Forecasting Stock Prices from the Limit Order Book Using Convolutional Neural Networks," 2017 IEEE 19th Conference on Business Informatics (CBI), Thessaloniki, 2017, pp. 7-12.

[5] T. Guo, J. Dong, H. Li and Y. Gao, "Simple convolutional neural network on image classification," 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(, Beijing, 2017, pp. 721-724.