Rain forecast problem based on weather characteristics

Carolina E. Machado
Depto. de Ciência da Computação
Universidade de Brasília
Brasília, DF, Brazil
carolina.machado@aluno.unb.br

***Abstract*—This article presents a comparison between two datasets using different feature extraction methods, such as Word2Vec and One-Hot-Encoding. The data is used to train a neural network and evaluate the results, comparing the two different approaches.**

***Index Terms*—Word2Vec, One-Hot-Encoding, neural networks**

I. Introduction

The reliability of weather and climate forecast is related to the knowledge of several atmospheric conditions parameters needed to power the models atmospheric circulation computations. The Artificial satellites provide a lot of relevant information that is used in the parameterization of the physical processes that occur in the atmosphere.

In this paper, two different weather datasets will be analyzed using numericalization methods as One-hot-encoding and Word2Vec representations similarly to [1] and compared using measures to quantitatively assess classification performance. There are two different approaches in this analysis, one based on binary classification and other on regression, which will be discussed later on section IV.

This paper represents an ongoing work to later devise a framework for weather data analysis obtained from satellite and ground-based images. Specifically, describing the design of the method for ground-based cloud image classification and the comparative study using convolutional neural networks and vision transformer. [2].

This paper is organized as follows: Section II presents the related works regarding approaches for weather data analysis. Section III describes the tasks and the proposed framework. Section IV presents the preliminary results on quantitatively assess classification performance of two public available datasets. Section V provides the final considerations and discussions over the next steps.

II. Related works

Zhang et al. [2] proposes a new convolutional neural network model, called CloudNet, a new framework of CNNs, which can achieve more accurate ground-based meteorological cloud classification. This new proposal shows progress comparing to other existing ground-based cloud databases as it is a more discriminative and comprehensive data set. The optimized model proposed consists of five convolutional layers and two fully connected layers. This model architecture suits well for distinguishing the different categories of cloud images and also to learn features of cloud representations. To evaluate the CloudNet model objectively, the CloudNet configuration was trained from the CCSN data set and SWIMCAT database as the pre-trained model. The results shows that the performance assessment of cloud classification of the CCSN database achieved a near-perfect classification accuracy for most categories. As for the SWIMCAT data set, the misclassification of the veil background and clear sky causes the poor performance of veil category, however the overall classification accuracy of CloudNet still outperforms the other traditional classification methods.

Isnain et al. [1] describes a study on detecting hate speech or not hate Indonesian speech tweets by using the of Bidirectional Long Short Term Memory method and the word2vec feature extraction method. A comparison between word2vec and one-hot-encoding is also made to determine the best models.

Shi et al. [3] proposes an algorithm using deep convolutional neural networks - a method proposed by Simonyan and Zisserman [4] for large-scale image recognition by adding more convolutional layers. The deep CNN model is used in this letter to extract feature representation for cloud images. The SWIMCAT Database and Kiel Database were used for cloud classification. After a series of experiments with these databases using deep convolutional activations-based features (DCAF) for ground-based cloud classification, the results show that the performance could be further improved after fine-tuning the pretrained models with cloud images and that the DCAF outperforms other traditional methods, further demonstrating the major superiority of learned DCAF over hand-crafted features for cloud classification.

III. Proposed method

The proposed method is divided into the following steps: Dataset definition, preprocessing, feature extraction, definition of the Neural Network model and performance comparisons and validation.

A. Dataset definition

The two datasets considered for this study represents characteristics about the weather, more specific about the rain type of precipitation.

I) *US Weather Events (2016 - 2021) dataset*: This is a countrywide weather events dataset that includes 7.5 million events, and covers 49 states of the United States. It presents 25424 instances, however, in this study, only 10000 instances were used. It is described by 10 attributes, but only 6 were used, according to table I

II) *Weather Dataset*: This dataset presents 49165 instances, however, as the previous one, only 10000 were used. It has

Attribute	Type
Precipitation(in)	Nominal
LocationLat	Number
LocationLng	Number
SeverityType	Nominal
StartDate	Datetime
StartTime	Datetime
EndDate	Datetime
EndTime	Datetime

TABLE I
US WEATHER EVENTS (2016 - 2021) DATASET ATTRIBUTES

Attribute	Type
Summary	Nominal
Precip Type	Nominal
Temperature (C)	Number
Apparent Temperature (C)	Number
Humidity	Number
Wind Speed (km/h)	Number
Wind Bearing (degrees)	Number
Visibility (km)	Number
Loud Cover	Number
Pressure (millibars)	Number
Date	Datetime
Time	Datetime

TABLE II
WEATHER DATASET ATTRIBUTES

similar attributes as the last one for comparison, according to table II

B. Preprocessing

During the process of working with these datasets, some preprocessing phases were needed, as deleting some attributes that weren't useful and merging some attributes into one more significant. It was also needed to remove instances with absent/null values.

C. Feature extraction

Different techniques were needed to treat the nominal and numerical attributes. For the numerical attributes, the MinMaxScaler normalization. Normalization is a technique to ensure that all data in the database have a similar range [5]. Eq. 1 and Eq.2 indicate the method of MinMaxScaler normalization.

$$X = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

$$X_{scaled} = X_{std} \times (X_{max} - X_{min}) + X_{min} \tag{2}$$

As for the nominal attributes, different approaches were used to obtain a structured representation of the texts. In this regard, Word2Vec and One-hot-Encoding were chosen.

Word2Vec, proposed and supported by Google, is not an individual algorithm, but it consists of two learning models, Continuous Bag of Words (CBOW) and Skip-gram [6]. By feeding text data into one of learning models, Word2Vec outputs word vectors represented as a large piece of text or even the entire article.

Similarly, One-hot-encoding is a Deep-Learned Embedding Technique for Categorical Features Encoding. One-hot encoding leads to very high dimensional vector representations, raising memory and computability concerns for machine learning models. [7]

After extracting this features, these two new datasets created are concatenated into one so that the model can use it for training.

D. Definition of the Neural Network model

The neural network used is architected as follows:

- *Dense layer*
- *Dense layer*
- *Dense layer*

The first layer includes an `input_dim` depending on the number of columns the dataset resulted from the merge mentioned before have and sets activation as "relu". The second layer consists of another Dense layer similar to the one before. Finally we have the output layer, which is also a Dense layer. It is defined by the type of function called *Linear* for the first dataset and *Softmax* for the second dataset.

It is also valid to compare the two different loss functions used on each model.

The Categorical Crossentropy function is used when we have a multi-class classification task. It is important to have the same number of output nodes as the classes and the final layer output should be passed through a softmax activation so that each node output a probability value between (0–1).

The Mean Squared Error function loss will be addressed later on section III-E.

E. Performance comparisons and validation

The evaluation Metrics used for comparisons were accuracy and Mean Squared Error. These results depends on how well the machine learning methods can predict the expected outputs.

Accuracy is a metric that generally describes how the model performs across all classes. It defines the rate of number of correct predictions and the total number of predictions. However, the accuracy may be a deceptive method, depending on how balanced the predicted classes are. 3

$$Accuracy = \frac{Numberofcorrectpredictions}{Totalnumberofpredictions} \tag{3}$$

Mean Squared Error (MSE) loss is used for regression tasks. As the name suggests, this loss is calculated by taking the mean of squared differences between actual(target) and predicted values.

As represented on tables III and IV, the accuracy for dataset I was higher than II. Nonetheless, the Mean Squared Error was lower for dataset II, more specific using the One-Hot-Encoding representation, which achieved the better result.

IV. EXPERIMENTAL RESULTS

In this section the experiment results are evaluated to make a comparison between them. The models were trained using the standardized parameters:

- Learning rate: 0.00002
- Epochs = 10; 30

There were two kinds of problems: A binary classification to know if there was precipitation or not and a problem of regression, to know the quantity of precipitation.

Binary classification is the process of classifying given document/account on the basis of predefined classes [8]. In

Representation	Dataset	Accuracy
One-Hot-Encoding	I	0.4247
Word2Vec	I	0.4500
One-Hot-Encoding	II	0.0010
Word2Vec	II	0.0010

TABLE III
COMPARISON BETWEEN DIFFERENT DATASETS ACCORDING TO ACCURACY

Representation	Dataset	Mean Squared Error
One-Hot-Encoding	I	0.1971
Word2Vec	I	0.1691
One-Hot-Encoding	II	0.0278
Word2Vec	II	0.2834

TABLE IV
COMPARISON BETWEEN DIFFERENT DATASETS ACCORDING TO MEAN SQUARED ERROR

this case, it is being used text categorization, which is also performed with binary classification.

The linear regression is one of the simplest and most common machine learning algorithms. It is a mathematical approach used to perform predictive analysis. [9]

The results in general were not satisfying, as the architecture and hyperparameters need to be optimized.

A multilayer Perceptron is one alternative of Perceptron model that can be used to optimize results. It has one or more hidden layers between its input and output layers, the neurons are organized in layers, the connections are always directed from lower layers to upper layers, the neurons in the same layer are not interconnected see Fig. 1.

The neurons number in the input layer equal to the number of measurement for the pattern problem and the neurons number in the output layer equal to the number of class, for the choice of layers number and neurons in each layers and connections called architecture problem, our main objectives is to optimize it for suitable network with sufficient parameters and good generalisation for classification or regression task. [10]

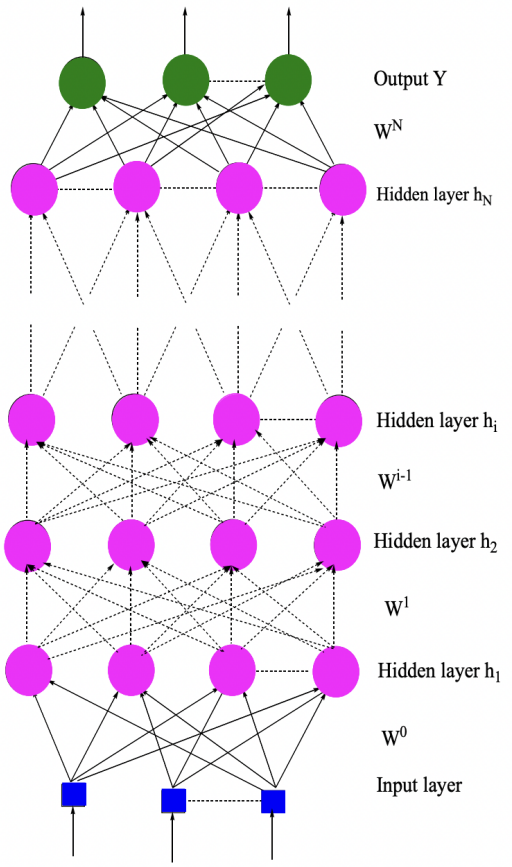

Fig. 1. Feed forward neural network structure.

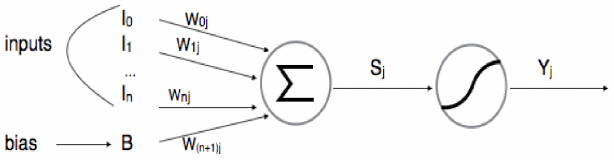

Fig. 2. Neuron Parameters.

V. CONCLUSION

This article evaluated the use of different models and embeddings to compare the efficiency of binary classification method and linear regression between two datasets. The method using a MinMaxScaler to deal with numerical attributes helped to facilitate the comparison between the Word2Vec and One-hot-Encoding techniques. As for the experimental results, the neural network along with Word2Vec presented better results for accuracy with dataset I, but the results for Mean Squared Error were better with One-Hot-Encoding for dataset II. Future work can be conducted by optimizing the architecture chosen for this experiment. Furthermore, other embedding techniques can be applied seeking for better results.

REFERENCES

[1] Y. S. Auliya Rahman Isnain, Agus Sihabuddin, "Bidirectional long short term memory method and word2vec extraction approach for hate speech detection," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 14, no. 2, 2020.

[2] J. Zhang, P. Liu, F. Zhang, and Q. Song, "Cloudnet: Ground-based cloud classification with deep convolutional neural network," *Geophysical Research Letters*, vol. 45, no. 16, pp. 8665–8672, 2018.

[3] C. Shi, C. Wang, Y. Wang, and B. Xiao, "Deep convolutional activations-based features for ground-based cloud classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 6, pp. 816–820, 2017.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[5] R. K. Deepa B, "Epileptic seizure detection using deep learning through min max scaler normalization," *International Journal of Health Sciences*, 2022.

[6] L. Ma and Y. Zhang, "Using word2vec to process big text data," in *2015 IEEE International Conference on Big Data (Big Data)*, 2015, pp. 2895–2897.

[7] M. K. Dahouda and I. Joe, "A deep-learned embedding technique for categorical features encoding," *IEEE Access*, vol. 9, pp. 114 381–114 391, 2021.

[8] S. K. S. Roshan Kumari, "Machine learning: A review on binary classification," *International Journal of Computer Applications*, vol. 160, no. 7, 2017.

[9] A. M. A. Dastan Hussen Maulud, "A review on linear regression comprehensive in machine learning," *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140–147, 2020.

[10] Y. G. M. E. Hassan Ramchoun, Mohammed Amine Janati Idrissi, "Multilayer perceptron: Architecture optimization and training," *Inter-*

national Journal of Interactive Multimedia and Artificial Intelligence,
vol. 4, no. 1.