Data Domain Change and Feature Selection to Predict Cardiac Pathology with a 2D Clinical Dataset and Convolutional Neural Networks (Student Abstract)

Mário Serra Neto,[1] Marco Mollinetti,[2] Inês Dutra [1]

[1] University of Porto (FCUP), Porto, Portugal [2] University of Tsukuba, Tsukuba, Japan
mariotrsn@gmail.com, ines@dcc.fc.up.pt, mollinetti@syou.tsukuba.ac.jp

Abstract

This work discusses a strategy named Map, Optimize and Learn (MOL) which analyzes how to change the representation of samples of a 2D dataset to generate useful patterns for classification tasks using Convolutional Neural Networks (CNN) architectures. The strategy is applied to a real-world scenario of children and teenagers with cardiac pathology and compared against state of the art Machine Learning (ML) algorithms for 2D datasets. Preliminary results suggests that the strategy has potential to improve the prediction quality.

Introduction

Neural Network (NN) is a popular ML technique that has been widely modified to learn data in multiple domains, e.g. time series, images, videos. Convolutional Neural Network (CNN), a particular variety of an NN, outperformed the classic model (Nielsen 2015) in the majority of tasks, especially in the field of computer vision.

The work developed by Miranda et al. (2019) applied a CNN model to classify events in power systems. In the approach, time series data was transformed into the shape of an image, which allowed the CNN to achieve 100% accuracy when classifying distinct types of events. To extend the approach, we propose the Map-Learn-Optimize (MOL) pipeline. The mapping task in MOL performs a feature sorting task of the dataset that will point out the order that features are inserted in the image format, then comes the optimization that will return the best image shape and which features should be used, increasing the similarity among generated images. Finally, the learning step evaluates a CNN model built to acquire knowledge from the generated images. The main goal of the proposed method is to search for novel ways of making CNN work well for data whose original format is not an image and compare with other machine learning models that are not based on neural networks.

Cardiac Pathology Data

The dataset was collected at a hospital specialized in cardiovascular system located at the northeastern part of Brazil. It is comprised of 20 features from 17,874 anonymous patients. Data was analyzed through bivariate and multivariate analysis, where preprocessing was performed in order to follow medical domain standards (e.g. age ranges, pressure rangers, etc) applying: data transformation; cleaning; normalization; removal of irrelevant features, such as ID and features with more than 95% of missing values. The processed dataset has 9,484 samples (53% of the original dataset) and 13 features: Weight; Height; Body-Mass-Index (BMI); Age; Wrist state (WS); Blood Pressure (PPA); Second heart sound (B2); Heart murmur (HM); Cardiac frequency (CF); Disease History (DH); Gender; Visit Reason (VR); and History Emergency Level (HEL).

MOL Strategy

At the mapping phase of MOL, the features are sorted according to a criteria. Sorted features are used to determine the row-wise order that features are inserted in the image format, which differs from Miranda et al. (2019) that inserted time-series in a row-wise format.

The optimization phase of MOL is an optimization task with the objective to maximize the similarity among the generated images, which is done by minimizing the following problem,

$$\begin{aligned} \text{minimize} \quad & \frac{1}{N}\sum_{i=1}^{N-1}\frac{1}{SM}\sum_{j=1}^{SM}(X_{i,j}-X_{i+1,j})^2 \\ \text{subject to} \quad & 2 \leq \hat{x_1}\hat{x_2} \leq M, \\ & \hat{x_1}\hat{x_2} = \sum_{i=3}^{M}\hat{x_i}. \end{aligned} \tag{1}$$

where *X* represents the dataset inputs features; *N* and *M* are respectively the number of instances and features found in the dataset; $\hat{x}$ is the solution for the problem with dimension 2+*M*; *SM* is the number of features of a feature subset *S*, where $S \subseteq D$. The first two variables of $\hat{x}$ are the image shape, while the rest are binary indicating if the feature will be used in the feature subset. The objective function attempts to minimize the mean squared error, subject to constraints that guarantee that the number of selected features will fit the image shape.

The learning phase of MOL consists into the train/test split, where the train set is used by the CNN algorithm un-

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

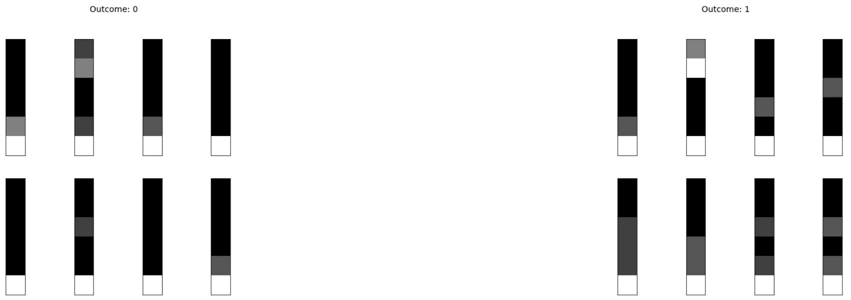

Figure 1: Absence and presence of cardiac pathology represented as image

der a validation strategy, generating a preliminary model that gives predictions and is evaluated by ML metrics.

Preliminary Experiment

We conducted an experiment using MOL to predict cardiac pathology in children and teenagers with the following experiment parameters: the map phase applied an Euclidean distance based strategy, e.g. features with the least distance between them were the first to be inserted; the optimization phase used Evolutionary Particle Swarm Optimization (EPSO) (Miranda and Fonseca 2002) with τ equal to 0.8, communication probability of 0.9, 50 particles and 500 function evaluations. The learn phase applied a CNN with 1 convolutional layer with 32 filters, 1 max pooling layer and a hidden layer with 8 neurons (mean of input and output).

The approach was compared against a Wrapper feature selection strategy (Chandrashekar and Sahin 2014) applied to a shallow neural network with same hidden neurons, trained by the Adaptive Moment estimation (ADAM) with learning rate equal to 0.001 and a Logistic Regression (LR) trained by Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm with l2 penalty. To handle the unbalanced dataset, the training phase was performed under a stratified 10-fold cross-validation, where the algorithms were evaluated through the mean of balanced accuracy obtained at each fold.

Preliminary Results

Sorting phase based on distance produced the following order for the CP dataset: Weight; WS; CF; B2; HEL; HM; Gender; PPA; DH; Age, Height, VR; and BMI. EPSO returned a solution where f= 0.43933, the image shape was equal to 6x1 represented by the features: B2; HEL; Heart murmur; Gender; PPA and Visit Reason. A sample of generated images using the approach is presented in Figure 1, where the outcome is the presence or absence of the disease. The sample shows that instances with absence outcome have more darker colors assigned to it, while instances with the presence of the pathology have lighter colors on the image. It is worth mentioning that the complete set of figures might present other patterns that can be easily detected by the CNN and not by human eye.

The results of training and testing can be found in Table 1. Preliminary results show that the approach achieved better results at both training and test phase, achieving the best result followed by the LR and ANN. A statistical analysis of the result using the Wilcoxon signed rank test suggests that the CNN approach had significance when compared to the others ($p < 0.05$), corroborating the fact that it is worthwhile to further study the MOL strategy in order to become a rival for state of the art approaches for 2D datasets.

Algorithm	Train	Test
CNN (MOL images)	0.936 ± 0.007	0.925
ANN (Original data)	0.914 ± 0.007	0.904
LR (Original data)	0.894 ± 0.069	0.872

Table 1: Results of the train/test phases

Final Remarks

This work investigated the MOL strategy to predict cardiac pathology in children and teenagers. The strategy is a process that applies data domain and feature selection to enable the application of CNN's to 2D datasets. Results indicated that MOL reached competitive results to predict cardiac pathology and might be promising in other areas of knowledge. The strategy will be further analyzed with the purpose to insert new map strategies, novel optimization algorithms, hyperparameter tuning and boosting.

References

Chandrashekar, G.; and Sahin, F. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* 40(1): 16–28.

Miranda, V.; Cardoso, P. A.; Bessa, R. J.; and Decker, I. 2019. Through the looking glass: Seeing events in power systems dynamics. *International Journal of Electrical Power & Energy Systems* 106: 411–419.

Miranda, V.; and Fonseca, N. 2002. EPSO-evolutionary particle swarm optimization, a new algorithm with applications in power systems. In *IEEE/PES Transmission and Distribution Conference and Exhibition*, volume 2, 745–750. IEEE.

Nielsen, M. A. 2015. *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, USA:.