ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Evaluating Curriculum Learning Strategies for Pancreatic Cancer Prediction

David Vázquez-Lema and Elena Hernández-Pereira and Eduardo Mosqueira-Rey *

Universidade da Coruña (CITIC) Campus de Elviña, A Coruña 15071, Spain

Abstract. In this work we applied Curriculum Learning (CL) to evaluate the performance of a machine learning (ML) model for pancreatic cancer prediction. As the dataset required it, we applied missing value imputation and data augmentation techniques. We compare different curriculum configurations in terms of pacing functions and we perform different experiments concluding that CL helps to train the ML model. Nevertheless, not all the configurations behave in the same way, and the best results were obtained when organising the curriculum in increasing levels of difficulty following exponential pacing.

1 Introduction

Pancreatic cancer is a very aggressive disease in which the identification of pancreatic tumours is a fundamental element in its diagnosis [1]. Among the different types of pancreatic cancer, the one with the highest incidence is Pancreatic Ductal Adenocarcinoma (PDAC) which if it is detected earlier, the survival rate can be greatly improved. Currently, a set of useful biomarkers are used to identify patients with some risk for developing PDAC [2].

Traditional ML algorithms randomly present all the training examples to the model, not paying attention to the different complexities of data examples and the learning status of the current model. Curriculum Learning (CL) is a training strategy that trains a ML model from easier data to complex data, which imitates the meaningful learning order in human curricula [3]. In contrast to CL, anti-CL selects the most difficult examples first and gradually exposes the model to easier ones. The power of introducing curriculum into ML depends on how the curriculum is designed for specific applications and datasets, but it has been demonstrated that CL helps to avoid bad local minima and leads to an improved generalisation [3].

The core of the CL strategy are a *scoring function* as the method to associate difficulty to examples, and a *pacing function* as the method to introduce examples to the training procedure. Among the scoring functions, one of the most used is *transfer teacher* where a stronger teacher model act as the teacher and measure the difficulty of training examples according to the teacher's performance on them [4]. The pacing function samples a batch of training data from the relatively easier examples (or more difficult in the case of anti-curriculum)

*This work has been supported by the State Research Agency of the Spanish Government, grant (PID2019-107194GB-I00/AEI/10.13039/501100011033). The authors wish to thank the funding received by Xunta de Galicia (grants ED431C 2022/44) and by CITIC (grant ED431G 2019/01 with European Regional Development Funds - ERDF).

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

and sends it to the model for training. With the progress of training, the pacing function will progress to sample from more harder (easier in anti-CL) data, until sampling from the whole training set.

In this work, we propose the application of supervised (automatic) CL for pancreatic cancer prediction. Our objective is to study different CL approaches in terms of pacing functions to demonstrate that the CL approach allows to improve the prediction of a base ML model, and to identify those examples that are difficult to learn. But one has to be careful with the choice of the pacing function, as not all offer the same results and benefits.

2 Material and methods

2.1 Dataset

To validate our proposal the *Urinary biomarkers for pancreatic cancer* [2] dataset was used. It gathers PDAC biomarkers samples collected from 590 patients before treatment: 183 were control samples (no disease), 208 benign disease samples and 199 PDAC samples (malign disease). Table 1 shows the dataset features used in the experiments. Demographics and clinical characteristics of patients can be consulted at [2].

Feature	Mean (std)	Missing rate
Age	59.1(13.1)	0%
Sex	F/M	0%
Blood plasma levels of monoclonal antibody (Serum CA19-9)	654(2.43k)	41%
Urinary biomarker of kidney function (Creatinine)	0.86(0.64)	0%
Lymphatic vessel endothelial hyaluronan receptor (LYVE-1)	3.06(3.44)	0%
Trefoil factor family peptide (TFF-1)	598(1.01k)	0%
Lithostathine-1-alpha protein (REG-1A)	735(1.47k)	48%
Regenerating family member 1 beta protein (REG-1B)	112(196)	0%

Table 1: Dataset description.

2.2 Data preprocessing and augmentation

As shown in Table 1, there are two features with a significant amount of missing values. Due to the limited data available, some data imputation methods were applied to determine and assign replacement values for missing data items. According to their degree of complexity, we have implemented two methods: a statistical method (mean) and a ML based method (k nearest neighbours, kNN) [5]. The first one is a method where any missing value is replaced by the mean of the observed values for that variable. The second one selects its k closest cases from the training cases with known values in the attributes to be imputed, such that they minimise some distance measure. Once the k nearest neighbours have

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

been found, a replacement value to substitute for the missing attribute value must be estimated. The replacement value is calculated using the mean and to determine the distance between training cases, the Euclidean measure was used.

Taking into account the variability of the features, a scaling method was applied and the sex categorical feature was transformed into a binary one. For the predicted feature (no disease, benign disease and malign disease) the one-hot encoding method was used.

As the quality of a ML model depends, among other things, on the quantity of the training data, data augmentation methods were tried. Taking into account the tabular nature of the dataset, SMOTE and CTGAN were used. SMOTE [6] is an oversampling approach in which the minority class is over-sampled by creating *synthetic* examples. These synthetic examples cause classifiers to create larger and less specific decision regions. This approach can improve the accuracy of classifiers for a minority class. Although the data set is balanced, we take the advantage of SMOTE to augment the examples. CTGAN [7] is a *GAN*-based model that generates synthetic tabular data. This implementation allows to define special relationships between columns, called *constraints*, that are used to improve the quality of the generated data by prohibiting certain combinations that may not exist in real data.

2.3 Curriculum design

As we pointed out previously, the two main elements in a supervised curriculum design are the scoring function and the pace function. The most general *transfer teacher* options are the loss based methods, which do not need any domain knowledge. Concretely, these methods take the example-wise losses calculated by a teacher model as the example difficulty and assume that the lower the loss, the easier the example [4]. After trying some transfer teacher options, we decided to adopt the bootstrapping strategy because of its simplicity and better proven results [8]. This strategy uses a teacher classifier with the same network structure as the student classifier, and pretrains it on the training dataset.

For the pacing function, we evaluate five different options:

- *One-Pass* (OP), that is a non-cumulative function where data is divided in equal size packages, in order to train every package a fixed amount of iterations and swap the package in use for the next one.
- *Baby-Step* (BS), that is similar to the OP function, but in this case it is cumulative, so the packages are merged every fixed amount of iterations.
- *Single-Step Pacing* (SSP), that starts training with a reduced amount of data a fixed amount of iterations, and then with the whole dataset.
- *Fixed Exponential Pacing* (FEP), that starts training with a reduced amount of data and exponentially increases it every fixed amount of iterations until all the data are used.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

- *Varied Exponential Pacing* (VEP), that is similar to FEP, but in this case the iterations are not fixed.

This CL design is contrasted against its *anti-curriculum* approach, i.e., the curriculum difficulties are complemented so that training samples follow the reverse order.

2.4 Base Machine Learning model and performance metrics

Adopting a bootstrapping strategy for the CL design means that a teacher ML model with the same structure as the student ML model should be designed. Our model, from now on the *base model*, is a deep network composed of a batch normalisation layer, two or three dense layers with ReLU or Tanh activation function, a dropout layer and an output layer with the *softmax* activation function. The model was built using Adam optimiser and a learning rate of 0.01.

In order to evaluate the quality and performance of the model, we use *F1-score* (F1) and *Macro-F1* (M-F1) metrics.

3 Experiments

We perform a comparative evaluation of the classification task with several experiments taking into account different network architectures, a specific data augmentation method and a specific imputation method, and different CL configurations. The experiments were performed taking the whole dataset and generate 5-fold cross validation sets in order to better estimate the performance of each model. In each experiment after the data augmentation method and imputation method selection, the deep model was optimised and the CL strategy was applied. From the experiments performed, we chose the most representatives (six):

- Experiments 1 and 2. In both experiments, SMOTE was used generating 34 new synthetic examples. Missing values were replaced by the mean in experiment 1 and by kNN in experiment 2. In experiment 1, the optimised model has three hidden layers with ReLU, and in experiment 2, three hidden layers with ReLU and Tanh intermixed.

- Experiments 3 and 4. In both experiments, no data augmentation was used and missing values were replaced by kNN. The optimised model has two hidden layers in experiment 3, and three hidden layers in experiment 4, both models with ReLU activation function.

- Experiment 5. Here, *CTGAN* was used to generate 200 new synthetic examples and missing values were replaced by kNN. The optimised model has three hidden layers with Tanh activation function.

- Experiment 6. Here, *CTGAN* was used to generate the same amount of examples as in experiments 1 and 2, and missing values were replaced by

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

kNN. The optimised model has two hidden layers with ReLU activation function.

4 Results

Table 2 shows the results for each of the six experiments in terms of F1 (for the three classes) and M-F1 measurements. This table includes the performance of the *base model* (no CL used) and the best of the different CL options for each experiment (we do not include all the results for space limitation). The Kruskal-Wallis test was applied to check if there are statistical differences among the test accuracies. The p-value obtained was $p < 0.001$ for a significance level of 95%. Therefore, the null hypothesis can clearly be rejected proving the statistical significance of the differences.

		F1 ND	**F1 BD**	**F1 MD**	**M-F1**
1	Base model	**0.670**	0.469	0.709	0.616
	CL with VEP	0.613	**0.587**	0.699	**0.633**
2	Base model	0.623	0.451	**0.747**	0.607
	CL with FEP	0.652	0.494	0.736	0.627
3	Base model	0.617	0.466	0.693	0.592
	CL with SSP	0.587	0.517	0.717	0.607
4	Base model	0.556	0.463	0.714	0.578
	CL with SSP	0.569	0.520	0.693	0.594
5	Base model	0.565	0.473	0.739	0.592
	CL with FEP	0.571	0.519	0.673	0.588
6	Base model	0.601	0.482	0.687	0.590
	CL with VEP	0.589	0.523	0.709	0.607

Table 2: Best results for each experiment for non disease (ND), benign disease (BD) and malign disease (MD). Best results are highlighted in bold.

The analysis of these results in terms of data augmentation, the replacement of missing values and CL design, leads us to the following statements:

- In general, data augmentation is beneficial for the performance of the model except when the number of synthetic examples is high respect to the dataset size. This fact can be observed at experiment 5. Comparing both SMOTE and CTGAN methods under the same conditions, SMOTE always outperforms CTGAN, helping the CL model to improve the performance of the classes where the base model has problems with.
- Taking into account the missing value replacement method, the mean method outperforms the kNN one, as experiments 1 and 2 demonstrate.
- From the CL strategy point of view, the curriculum strategy achieves the best results in terms of M-F1 and almost always outperforms the anti-curriculum strategy. Taking into account the pacing function, the One-

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Pass method is not useful in this classification task as it always had the worst results. Its M-F1 value for all the experiments does not exceed the value of 0.521. On the other hand, the exponential versions of the pacing function achieve the best results in terms of performance.

5 Conclusions

In this work we have proposed to apply CL to evaluate the performance of a ML model. It seems that this technique helps the ML model in its classification task, but it's really dependent on the selected CL configuration. In particular, the pacing function has a huge influence on the results. If the pacing function is changed, the results can dramatically change too. It is preferable to organise the training procedure presenting the examples increasing its difficulty in exponential steps, and to avoid non-cumulative methods (OP). Nevertheless, the dataset used is limited and it does not allow the CL strategy to improve significantly the performance of the base model. Since differentiating the easy examples from the difficult ones could be a complex task in this dataset, we envisage as future work to add an expert point of view (*Human-in-the-loop* [9]) to help develop a better scoring function that would improve the results.

References

[1] Melina Arnold et al. Progress in cancer survival, mortality, and incidence in seven high-income countries 1995â2014 (icbp survmark-2): a population-based study. *The Lancet Oncology*, 20(11):1493–1505, 2019.

[2] Silvana Debernardi et al. A combination of urinary biomarker panel and pancRISK score for earlier detection of pancreatic cancer: A case-control study. *PLOS Med*, 17(12), 2020.

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41â48, New York, NY, USA, 2009. Association for Computing Machinery.

[4] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2022.

[5] Pedro J García-Laencina, José-Luis Sancho-Gómez, Aníbal R Figueiras-Vidal, and Michel Verleysen. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing*, 72(7-9):1483–1493, 2009.

[6] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321â357, jun 2002.

[7] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[8] Himmet Toprak Kesgin and Mehmet Fatih Amasyali. Development and comparison of scoring functions in curriculum learning. In *2022 2nd International Conference on Computing and Machine Intelligence (ICMI)*. IEEE, apr 2022.

[9] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56:3005–3054, 2023.