Synthetic Data Augmentation of Cycling Sport Training Datasets

Iztok Fister, Grega Vrbančič, Vili Podgorelec, and Iztok Fister Jr.(✉)

University of Maribor, Koroška Ul. 43, 2000 Maribor, Slovenia
iztok.fister1@um.si

Abstract. Planning sport sessions automatically is becoming a very important aspect of improving an athlete's fitness. So far, many Artificial Intelligence methods have been proposed for planning sport training sessions. These methods depend largely on test data, where Machine Learning models are built, yet evaluated later. However, one of the biggest concerns of Machine Learning is dealing with data that are not present in the training dataset, but are unavoidable for predicting the further improvements. This also represents a bottleneck in the domain of Sport Training, where algorithms can hardly predict the future training sessions that are compiled with the attributes of features presented in a training dataset. Usually, this results in an under-trained trainee. In this paper we look on this problem, and propose a novel method for synthetic data augmentation applied on the original dataset.

Keywords: Synthetic data augmentation · Cycling sport training · TRIMP

1 Introduction

The performance of Machine Learning (ML) approaches, methods and techniques is largely dependent on the characteristics of the target dataset, which consists of numerous training samples. Such datasets, that contain a higher number of samples, which are also more diverse, in general, offer more expressive power to the learning algorithms applied in various ML models, e.g. classification, regression. When the datasets are not too diverse or the samples in datasets are not representing all possible features in the real world, algorithms may suffer in their ability to deal with their tasks, and, consequently, cannot deliver the optimal outcome for a particular classification/regression problem. That behavior can also be pointed out as a drawback of ML [14].

The lack of diverse datasets and small number of samples in datasets in the field of Data Science is, nowadays, commonly addressed using various data augmentation methods. These methods are able to increase the amount of data by either adding slightly modified copies of already existing data, or creating new synthetic data from existing ones. Beside the primary goal, the data augmentation techniques also attempt to increase the generalization abilities of ML models by reducing the overfitting and expanding the decision boundary of the

P. Vasant et al. (Eds.): ICO 2021, LNNS 371, pp. 65–74, 2022.
https://doi.org/10.1007/978-3-030-93247-3_7

model [12,21]. There are many techniques for increasing the generalization of ML models, such as dropout [23], batch normalization [11] or transfer learning [19,25]. However, the data augmentation techniques address the mentioned problem from the root, which is the training dataset. While for certain ML problems, such as image classification, the established practices of data augmentation exist, this is not the case when dealing with time-series datasets or sports datasets, which are commonly in structured form, with features extracted from time-series datasets. With this specific of sports datasets in mind, the common existing approaches to data augmentation would not result in an expected performance improvement of the predictive ML model. Therefore, the need arises for a domain-specific method for synthetic data augmentation.

Sport is becoming a very interesting venue for ML researchers, where they are confronted with building ML models based on past training data in order to plan future sport training sessions for a particular person. Several solutions exist in the literature [6,20,22]. Artificial Sport Trainer (AST) [5] is an example of a more complex solution for assisting athletes with automatic planning of sport training sessions. Besides the planning, AST is also able to cover the other phases of sport training, i.e. realization, control, and evaluation.

The main weakness of the AST is that it is unable to identify training sessions with intensities beyond those presented in an archive. This paper introduces a synthetic data augmentation method for generating uncommon sport training sessions. This means that the method is appropriate to prescribe training sessions which ensure that athletes in training will improve their fitness further. The method consists of more steps, in which features need to be identified first. Then, the interdependence among these are captured using the new Training Stress Measure (TSM). This metric serves for recognizing the most intensive training sessions, serving as a basis for generation of the uncommon training session using synthetic data augmentation. Finally, the set of uncommon training sessions enrich the existing archive.

The proposed method was applied on an archive of sport training sessions generated for an amateur cyclist by the SportyDataGen generator [7]. The results showed that the method can also be used for enriching the archive of the sport training sessions in practice.

The contributions of this paper are:

- to propose a method to augment the training dataset of existing sport activities with synthetic data augmentation,
- to evaluate the method on data suitable for an amateur cyclist.

2 Problem Statement

Two subjects are important for understanding the subjects that follow:

- data augmentation,
- basics of cycling training.

The former highlights the principles of data augmentation, with emphasis on the synthetic data augmentation, while the latter reveals the basics of cycling sport training.

2.1 Data Augmentation

In general, the term data augmentation covers various strategies, approaches and techniques which are trying to increase the size of the training dataset artificially, or to make the utilized dataset more diverse in order to represent the real life distribution of the training data better.

For the purpose of tackling image classification tasks, the data augmentation is an already established practice, especially when utilizing deep Convolutional Neural Networks (CNN) [17]. Dating back to 2009, when one of the first well-known CNN architectures, AlexNet [16], was presented, which was utilizing techniques for cropping, mirroring, and color augmentation of training datasets in order to improve the generalization capabilities of the trained ML model, as well as reducing the problem of overfitting [12]. The approaches and techniques for augmentation of images could be divided in two categories: classical image data augmentation also known as basic data augmentation and deep learning data augmentation [15]. The first group of approaches consists primarily of techniques which are manipulating the existing training dataset with geometric transformations and photometric shifting. The geometric transformation techniques include flipping images horizontally or vertically, rotating, shearing, cropping, and translating the images, while the photometric techniques cover color space shifting, adding image filters and introducing the noise to the images [15]. The more advanced approach to image data augmentation is with utilization of deep learning. Such approaches could be split into three groups: generative adversarial networks [10], neural style transfer [13], and meta metric learning [8]. The last group includes the techniques which are utilizing the concept of utilization of a neural network to optimize neural networks. Some of the representatives of this kind of data augmentation techniques are neural augmentation [24], auto augmentation [4], and smart augmentation [18].

While the approaches for the image data augmentation are well explored, and are, in general, a part of a standard procedure, this is not yet the case for the augmentation for time-series datasets. However, the researches in this field have been gaining new momentum in recent years. Similar to the image data augmentation techniques, some of the common approaches for augmentation of time-series datasets are based on random transformations of the training data, such as adding noise, scaling or cropping. However, the problem with utilization of such approaches when dealing with time-series datasets is that there is a diverse amount of time-series which each have different features, and not every transformation is therefore suitable for application on every dataset. Based on those specifics of time-series datasets, alternative approaches were developed, such as synthesis of time-series. The group of synthesis based approaches includes a number of different methods and techniques, from pattern mixing, generative models, and decomposition methods [12].

While the presented approaches are proven to work well on specific target tasks, the sports data are somewhat specific in terms of data augmentation. Since, in general, such datasets are commonly derived from the recorded training sessions of different athletes, which are in fact time-series datasets, one could try

to utilize the known techniques for time-series dataset augmentation and extract the features afterwards. However, such an approach would not be able to provide us with the data inferred from the training sessions beyond those presented in the original dataset. In other words, with such an approach, we could not augment the original dataset in such a manner that we would obtain features for more uncommon sport training sessions, but could only obtain a slightly more diverse variation of already present training sessions.

2.2 Basics of Cycling Training

The goal of sport training is to achieve a competitive performance for an athlete that is typically tested in competitions. The process is directed by some principles, where the most important are the following two: The principle of progressive load and the principle of reversibility [9]. According to the first, only a steadily increasing the amount of training can lead to an increase in an athlete's fitness, while, according to the second, the fitness is lost by the athlete's inactivity, rest, or recovery. The athlete's body senses the increased amount of training as a physical stress. When this stress is not too high, fatigues are experienced by the athlete after the hard training. These can be overcome by introducing the rest phase into the process or performing the less intensive workouts. On the other hand, if the stress is too high, we can talk about overtraining. Consequently, this demands the recovery phase, in which the athlete usually needs the help of medical services to recover.

The amount of training typically prescribed by a sport trainer, is determined by a training load. The training load specifies the volume of training and the intensity at which it must be realized. Thus, the volume is simply a product of duration and frequency. With these concepts in mind, the training plan can be expressed mathematically as a triple:

$$\text{Training plan} = \langle \text{Duration, Intensity, Frequency} \rangle. \tag{1}$$

Duration denotes the amount of the training. Typically, the cycling training is specified by a duration in minutes. Intensity can be determined by various measurements, but mostly the following two are accepted in cycling: Heart Rate (HR) and Power (PWR). Although the PWR presents accurate intensity measurements [1], this study focuses on the HR due to a lack of experimental data including it. Based on the so-called Functional Threshold Heart Rate (FTHR), specific for each individual athlete, the HR intensity zones are defined that help trainers to simplify the planning process. Let us mention that the FTHR is an approximation of the maximum HR (max HR) denoting an Anaerobic Threshold (AnT), where a lactate drop in the blood started to accumulate in the body due to insufficient oxygen intake [3]. The frequency is a response to the question of how many times the training is performed during the same period.

An example of the HR intensity zones suitable for an athlete with FTHR = 180 (40 year old athlete) is presented in Table 1, from which it can be seen that there are seven HR zones with their corresponding names. The column

Table 1. HR zones suitable for a 40-year-old athlete with FTHR = 180.

HR zone	Name	% FTHR	FTHR
Zone-1	Active recovery	<60	<108
Zone-2	Aerobic endurance	60–70	108–126
Zone-3	Tempo	70–80	126–144
Zone-4	Lactate threshold	80–90	144–162
Zone-5	VO_2max	90–100	162–180
Zone-6	Anaerobic capacity	≥100	≥180

"% FTHR" denotes the bounds of the particular HR zone interval expressed as a percentage of FTHR, and the column "FTHR" their absolute values.

3 Proposed Method

When the AST generates the training plan based on the existing training sessions, it is confronted with the same problem as the other ML methods, i.e., how to handle objects of qualities beyond those presented in the testing database. Moreover, in the theory of Sport Training, this limitation even violates the training principle of progressive overload, stating that the progressively harder training plan needs to be performed by athletes in training to increase their fitness.

As a result, a data augmentation method is proposed in the paper to overcome the problem. The method consists of the following steps:

- identification of features in an archive of sport activities,
- identification of interdependence among features,
- identification of the most intensive training sessions,
- generation of uncommon training sessions based on synthetic data,
- data enrichment of the existing archive with the uncommon training sessions.

An archive of sport activities needs to be collected in the first step. Usually, these activities are obtained from athletes worn mobile devices capable of monitoring athletes' performances during the realization of training sessions. Unfortunately, collecting the sport activities in this way is slightly complicated, because obtaining the data needs the permission of the athlete. On the other hand, the number of sport activities is limited by the number of realized sport sessions by the athlete. In line with this, an online generator of endurance sports activity collections has been proposed by Fister et al. [7] that was also used in our study. The generator is capable of generating the collection of a specific number of fields containing features for various sport disciplines and profiles of athletes. Here, we are focused on endurance cycling, and, thus, manipulate with the features presented in Table 2. As can be seen from the Table, each sport activity is identified by its identifier and six features representing training load indicators achieved by an athlete during the sport training session. Interestingly, all attributes are either real or integer numbers.

Table 2. Features and the corresponding domains of attributes.

Nr.	Feature name	Abbreviation	Attribute's domain
1	Sport activity identifier	*ID*	$ID \in \mathcal{N}$
2	Duration of the sport activity	*Duration*	$Duration \in \mathcal{R}$
3	Distance of the sport activity	*Distance*	$Distance \in \mathcal{R}$
4	Average HR	$\overline{HR}$	$\overline{HR} \in \mathcal{N}$
5	Burned calories	*Calories*	$Calories \in \mathcal{N}$
6	Average altitude	$\overline{Alt}$	$\overline{Alt} \in \mathcal{R}$
7	Maximum altitude	$Alt_{\max}$	$Alt_{\max} \in \mathcal{R}$

The observed features highlight performances of athletes achieved during the realization of the training session. The simplest measure for estimation of physical stress was TRaining IMPulse (TRIMP), proposed by Banister [2]. The TRIMP is expressed as:

$$\text{TRIMP} = Duration \cdot \overline{HR}, \tag{2}$$

and highlights the relationship between the duration and average HR. If we want to take all relations among the six observed features into account, the new measure is necessary. Indeed, we propose the so-called Training Stress Measure (TSM) defined as follows:

$$\text{TSM} = K \cdot \frac{Distance \cdot \overline{HR} \cdot Calories}{Duration}, \tag{3}$$

where K is expressed using the following equation:

$$K = \frac{\overline{Alt}}{Alt_{\max}}. \tag{4}$$

As can be seen from Eq. (3), the TSM is an extension of the already mentioned measure TRIMP, and reflects the relationships between all features except *Calories*. This feature is connected strongly with power meters. In the case of using the HR trackers (as in our study), this value is only estimated, and therefore can be omitted.

The motivation behind the generation of the uncommon training sessions is to take the potential candidate from the set of uncommon training sessions and ride the same course more intensively (i.e., with a higher average HR). In line with this, only three load indicators remain important for generation of uncommon training sessions, i.e., *Duration*, *Distance* and $\overline{HR}$. Moreover, the first two indicators refer to the speed of riding, in other words: $Speed = Duration/Distance$, while the third determines the intensity of the training session. Increasing this indicator means raising the intensity, and, indirectly, also the speed. How the speed will be increased we cannot express explicitly, due to the psycho-physical characteristics of an athlete, but it must be predicted.

In this study, the $\overline{HR}$ is increased by ΔHR in order to change the potential candidate to a really uncommon training session. The speed is predicted as follows: The linear regression between *Speed* and modified heart rate $\overline{HR} + \Delta HR$ is calculated. As a result, a regression line $Y = a \cdot X + b$ is calculated on the potential candidate, and moving the regression line upward for offset proportional to increasing the average HR serves us as a tool for predicting the new *Speed* value. After obtaining this value, the new value of TSM is calculated according to Eq. (3).

The M uncommon training sessions must be included into a collection of training session, thus, enriching it with more intensive training sessions.

4 Experimental Results

The goal of our experimental work was to show that uncommon sport activities can improve the performance of an athlete using synthetic data augmentation, and thus obey the first principle of the sport training. The experiments followed the steps recommended by the proposed method for synthetic data augmentation.

At first, a collection of $N = 500$ sport activities was generated by SportyDataGen [7]. Then, a set of potential uncommon sport activities was identified, where the $M = 10$ of the most intensive training sessions were extracted. This set served as potential candidates from which the uncommon training sessions were generated by increasing the intensity of the candidate training sessions by $\Delta HR = 1$. Thus, it is assumed that the small stepwise increasing of the intensity is enough to improve the fitness of an athlete, measured by TSM, significantly. Finally, the set of uncommon training sessions was enriched with the existing collection and the new training plan was proposed by the CI algorithm for planning training sessions from sport activities.

Table 3. Characteristics of a collection of training sessions ($FTHR = 180$).

HR zone	*FTHR*	Effective HR	Sessions
Zone-1	<108	72–107	39
Zone-2	108–126	108–125	229
Zone-3	126–144	126–144	193
Zone-4	144–162	144–151	39
Zone-5	162–180	n/a	0
Zone-6	>180	n/a	0
Avg./Total		**111.50**	**500**

The characteristics of the collection are illustrated in Table 3, from which it can be seen that it's about a 40 year old amateur athlete, with the majority of sport activities realized in intensity Zone-2 and Zone-3 (i.e., aerobic endurance and tempo). The 39 sport activities belong to intensity Zone-1, and the same number of activities to intensity Zone-4. These characteristics show that the cyclist was prepared primarily for endurance events.

A set of the $M = 10$ of the most intensive training sessions from a collection of sport activities are presented in Table 5, from which it can be seen that only two training sessions from the set belong to the most intensity Zone-4. These sessions are presented in bold in the Table 4.

Table 4. The most intensive training sessions according to the TSM ($M = 10$).

ID	HR zone	*Duration*	*Distance*	$\overline{HR}$	*Calories*	*Alt*	$Alt_{\max}$	*TSM*
427	3	62.43	39.97	136	610	542.36	543.4	5214.03
347	3	62.87	37.94	143	549	842.58	844.2	5168.71
124	3	70.02	42.49	138	423	770.14	772.6	5008.55
183	4	57.52	34.41	**144**	700	616.91	642.6	4963.02
338	3	70.12	41.01	136	550	308.24	308.8	4763.54
301	3	65.22	38.64	134	614	178.64	182.8	4655.43
462	4	59.2	34.43	**148**	822	202.65	225.4	4643.59
401	3	66.45	39.72	129	622	354.31	356.6	4596.28
500	3	59.92	36.73	130	561	434.58	456.6	4550.31
459	3	62.8	37.2	128	501	1535.69	1538.4	4541.82
Total								**48105.28**

The set of the most intensive sport training sessions constituted a set of potential candidates, from which a set of uncommon training sessions was generated by increasing the intensity of the candidate training sessions $\overline{HR}$ for $\Delta HR = 1$ (Table 5). As can be seen from this Table, the motivation behind

Table 5. The generated set of uncommon training sessions ($\Delta HR = 1$).

ID	*Duration*′	*Distance*	$\overline{HR}'$	*Calories*	*Alt*	$Alt_{\max}$	TSM’	ΔTSM
501	66.51	39.97	137	610	542.36	543.40	4930.18	−283.85
502	63.48	37.94	**144**	549	842.58	844.20	5154.50	−14.21
503	70.81	42.49	139	423	770.16	772.60	4988.29	−20.26
504	57.62	34.41	**145**	700	616.91	642.60	4988.54	25.52
505	68.24	41.01	137	550	308.24	308.80	4930.71	167.17
506	64.21	38.64	135	614	178.64	182.80	4764.02	108.59
507	57.83	34.43	**149**	822	202.65	225.40	4786.10	142.51
508	65.74	39.72	130	622	354.31	356.60	4681.91	85.63
509	60.84	36.73	131	561	434.58	456.60	4515.98	−34.33
510	61.54	37.20	129	501	1535.69	1538.40	4671.21	129.39
Total							**48411.43**	**176.77**

increasing the intensity of potential sport activities was to realize the particular activity in a higher intensity zone. In line with this, the candidate with $ID = 347$ in Zone-3 moved to a set of the uncommon training session belonging to Zone-4.

In summary, increasing the intensity of candidate sport activities caused increasing of the total TSM by $\Delta\text{TSM} = 176.77$ units.

5 Conclusion

One of the major bottlenecks of the ML methods is the fact that the data necessary for predicting the further improvements are not presented in the test datasets. In line with this, data augmentation methods have been developed that increase the amount of data by creating new synthetic data from existing ones.

In our study, the synthetic data augmentation was applied to ARM, that is one of the famous ML methods. The ARM method is a part of the AST, capable of analyzing existing sport training activities. These sport activities serve as a basis for planning sport training sessions in various sport disciplines. We focused on synthetic data augmentation in cycling sport, while the results showed its huge potential for use in practice.

There are several directions for future development of the proposed method, such as for instance: (1) Using the synthetic data augmentation in other sports disciplines, like running, triathlon, etc., (2) Widening the number of training load indicators, (3) Incorporating the power meter data into analysis, and (4) Inventing the new training stress measure, also able to express calories and power meter data.

References

1. Allen, H., Coggan, A.R., McGregor, S.: Training and Racing with a Power Meter, 3rd edn. VeloPress, Boulder (2019)
2. Banister, E.W.: Modeling elite athletic performance. Physiol. Test. Elite Athletes **347**, 403–422 (1991)
3. Clark, M.A., Lucett, S.C., Sutton, B.G.: NASM Essentials of Personal Fitness Training, 4th edn. Jones & Bartlett Learning, Burlington (2014)
4. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: AutoAugment: learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 113–123 (2019)
5. Fister, I., Fister Jr., I., Fister, D.: Computational Intelligence in Sports. ALO, vol. 22. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-03490-0
6. Fister, I., Rauter, S., Yang, X.-S., Ljubič, K., Fister Jr., I.: Planning the sports training sessions with the bat algorithm. Neurocomputing **149**, 993–1002 (2015)
7. Fister Jr., I., Vrbancic, G., Brezočnik, L., Podgorelec, V., Fister, I.: SportyDataGen: an online generator of endurance sports activity collections. In: CECIIS: Central European Conference on Information and Intelligent Systems, pp. 171–178. IEEE (2018)
8. Frans, K., Ho, J., Chen, X., Abbeel, P., Schulman, J.: Meta learning shared hierarchies. arXiv preprint arXiv:1710.09767 (2017)

9. Friel, J.: The Cyclist's Training Bible: The World's Most Comprehensive Training Guide, 5th edn. VeloPress, Boulder (2018)
10. Goodfellow, I.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
11. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456. PMLR (2015)
12. Iwana, B.K., Uchida, S.: An empirical survey of data augmentation for time series classification with neural networks. PLoS ONE **16**(7), e0254841 (2021)
13. Jing, Y., Yang, Y., Feng, Z., Ye, J., Yizhou, Yu., Song, M.: Neural style transfer: a review. IEEE Trans. Visual. Comput. Graph. **26**(11), 3365–3385 (2019)
14. Kauwe, S.K., Graser, J., Murdock, R., Sparks, T.D.: Can machine learning find extraordinary materials? Comput. Mater. Sci. **174**, 109498 (2020)
15. Khalifa, N.E., Loey, M., Mirjalili, S.: A comprehensive survey of recent trends in deep learning for digital images augmentation. Artif. Intell. Rev., 1–27 (2021)
16. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
17. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
18. Lemley, J., Bazrafkan, S., Corcoran, P.: Smart augmentation learning an optimal data augmentation strategy. IEEE Access **5**, 5858–5869 (2017)
19. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2009)
20. Rauter, S.: New approach for planning the mountain bike training with virtual coach (2018)
21. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 1–48 (2019)
22. Silacci, A., Taiar, R., Caon, M.: Towards an AI-based tailored training planning for road cyclists: a case study. Appl. Sci. **11**(1), 313 (2021)
23. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
24. Wang, J., Perez, L., et al.: The effectiveness of data augmentation in image classification using deep learning. Convolutional Neural Netw. Vis. Recogn. **11**, 1–8 (2017)
25. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. J. Big Data **3**(1), 1–40 (2016)