A SMOTE Extension for Balancing Multivariate Epilepsy-related Time Series Datasets

Enrique de la Cal[1], José R. Villar[1], Paula Vergara[1], Javier Sedano[2], and Álvaro Herrero[3]

[1] University of Oviedo, EIMEM, Independencia 13, Oviedo 33004, Spain
delacal@uniovi.es
[2] Instituto Tecnol'ogico de Castilla y León,
Polg. Ind. Villalonquejar, L'opez Bravo 70, Burgos 09001, Spain
[3] Department of Civil Engineering, University of Burgos, Spain ahcosio@ubu.es

Abstract. In some cases, big data bunches are in the form of Time Series (TS), where the occurrence of complex TS events are rarely presented. In this scenario, learning algorithms need to cope with the TS data balancing problem, which has been barely studied for TS datasets. This research addresses this issue, describing a very simple TS extension of the well-known SMOTE algorithm for balancing datasets. To validate the proposal, it is applied to a realistic dataset publicly available containing epilepsy-related TS. A study on the characteristics of the dataset before and after the performance of this TS balancing algorithm is performed, showing evidence on the requirements for the research on this topic, the energy efficiency of the algorithm and the TS generation process among them.

Keywords: Dataset balancing algorithms, SMOTE, Time Series

1 Introduction

In recent years, new technological challenges and opportunities are being discovered, such as Industry 4.0, Internet of the Things or e-Health, where vast amounts of data are produced and gathered. In some of the cases, these big data bunches are in the form of Time Series (TS). Such cases include the management of sensory systems located on wearable devices, as in the problems of human activity recognition and abnormal movement detection [1,2,3]. Furthermore, TS datasets have become multivariate datasets, which makes the data analysis even more complex.

In this context, when leaning models for the detection of some complex events, the problem of unbalanced data arises: there are many more TS segments belonging to normal class than those belonging to the abnormal class to be detected. For instance, in the problem of epilepsy seizure detection [3,4], the occurrence of a seizure might be once in a month or even less.

Most previous work on the dataset balancing problem is focused on classical datasets, where a sample includes an atomic value for each of the features.

These balancing techniques rely on oversampling the minority class (mC) or undersampling the majority classes (MC); however, as long as oversampling does not produce information losses, it is preferred over undersampling.

Some valid alternatives have also been published, coping with imbalanced problems specific algorithms [5], or proposing ensembles for the mC together with a kind of undersampling of the MC [6]. Examples of oversampling techniques include well-known algorithms such as SMOTE (Synthetic Minority Over-sampling Technique [7,8]), ADASYN (ADAptive SYNthetic Sampling, [9]) ADOMS (Adjusting the Direction Of the synthetic Minority clasS examples [10]) or SPIDER (Selective Preprocessing of Imbalanced Data [11]).

However, the problem of balancing TS datasets has received scant attention from the scientific community up to now. The published approaches focused on univariate TS problems [12,13,14,15], where the known data sequence labels are clearly biased to the MC. Therefore, the solutions rely on drawing new synthetic atomic values based on any of the above mentioned algorithms. On the other hand, Koknar et al [16] proposed the balancing of univariate TS based on suggesting ghost points. These ghosts points belong to the domain space of TS distances. With the associated distance matrix an SVM classifier is trained; allowing to generate a new TS and assigning it to a class. Different TS distance measurements were proposed, such as the Dynamic Time Warping (DTW). Besides, in a multivariate TS dataset problem, each sample in the TS dataset includes a TS for each feature. Moreover, the sample is assigned a class, but also a TS is attached as the labelling TS for that sample. From now on, we consider all the TS features from a sample with the same length and sampling frequency; however, the variability in these factors needs further study. This study addresses the multivariate TS datasets balancing problem, extending the well-known SMOTE algorithm to cope with multivariate TS. The experimentation will analyze the distortion in the dataset due to the inclusion of the new TS samples. This study is structured as follows. Next section outlines the SMOTE algorithm, while the design issues and possible solutions are given in Sect. 3. Experimentation and the discussion on the obtained results are coped in Sect. 4. Finally, the main conclusions are drawn in Sect. 5.

2 The SMOTE Algorithm

The SMOTE algorithm is an oversampling method [7] where new synthetic data are generated to balance a given dataset. In order to do that, each sample from the mC is randomly combined with each one of its nearest neighbors. This method assumes a two-class problem, however, it can be easily extended to multi-class problems [17].

The original SMOTE algorithm from the seminal paper [7] is shown in Algorithm 1. The parameters of this method include the number of nearest neighbors to be considered (k, a default value of $k = 5$ has been proposed), the number of samples belonging to the mC (T) and the number of synthetic samples to be generated for each original sample from the mC (N). This parameter N is

given as a percentage; values smaller than 100% reduces the original minority subset and produces a new dataset of the same size as the original. Whenever $N > 100\%$ means that $N/100$ synthetic samples are to be generated for each one of the samples from the mC.

Algorithm 1 The SMOTE original algorithm. Three parameters (T, N, k) are needed, as stated above.
SMOTE(T, N, k)

```
if N < 100 then
    Randomize the T minority class samples
    T = (N / 100) * T
    N = 100
end if
N = int( N / 100 )
numattrs = Number of attributes
Sample[][]: array for original minority class samples
newindex: counts the number of generated synthetic samples
Synthetic[][]= array for synthetic samples
for i = 1 : T do
    Compute the k nearest neighbors of sample i, saving the indexes in nnarray
    Populate(N, i, nnarray)
end for
function POPULATE(N, i, nnarray)
    while N ≠ 0 do
        Choose a random number nn in {1, k}
        for  attr = 1 : numattrs do
            dif = Sample[nnarray[nn]][attr]-Sample[i][attr]
            gap = random number in {0, 1}
            Synthetic[newindex][attr] = Sample[i][attr] + gap * dif
        end for
        newindex ++
        N = N - 1
    end while
end function
```

As it can be seen in Algorithm 1, SMOTE takes a sample and searches for some neighbors; each synthetic sample is generated as a random linear combination of the two considered samples. This method has been successfully tested on different domains; and plenty of different versions have been published [8]. Some improvements on the original algorithm include i) cleaning the new dataset of mC *Tomek links* producing the *SMOTE+Tomek links* and ii) cleaning the whole dataset of *Tomek links*, known as *SMOTE-ENN*. A Tomek link is a sample from one class that is included in the counterpart class space. Formally speaking, a pair of samples E_i and E_j -labelled with different classes- forms a Tomek link if there no exists a sample E_k such that $d(E_i, E_k) < d(E_i, E_j)$ or $d(E_j, E_k) < d(E_i, Ej)$,

with d being the distance function. However, these method are a sort of denoising stage, and that is the reason why they are not considered in this study.

3 Tackling the TS Balancing Problem through AVG_TS_SMOTE

At least two main concerns have to be solved in order to allow the SMOTE algorithm to cope with TS datasets. The first one is related to the method for choosing the parents TS samples to mate, while the second focuses on the generation of the new TS sample.

When choosing the two TS samples that will be used for generating the new TS offspring, TS grouping according to some measurement should be considered. The original SMOTE algorithm randomly selects the parents for mating among those belonging to the mC. However, different and more advanced solutions can be considered; for instance, the solution proposed in ADASYN [9], where the parents are randomly chosen according to the distribution of the size of the neighborhood, is totally valid as well. It seems, according to the published results for the different SMOTE-based flavors, that problem-oriented heuristics might be the best solution for each problem. An example of such heuristic can be grouping the TS samples for the mC using the mean value of the Phan et al distance [18]; afterwards, two different groups are randomly chosen; finally, one TS sample is chosen from each one of the two candidate groups. Nevertheless, the best performance of this distance measurement is obtained when the length of the TS is bounded to less than a relatively small value.

On the other hand, the generation of a new TS when oversampling is not a simple task: as long as multivariate TS are considered, the new TS sample will need a TS for each one of the available features. For each feature to be generated, a combination of the parents' same feature should be performed. Furthermore, the combination must be coherent for all the features considered as a single sample. Finally, the TS class has to be generated as well, which is much of a compromise. Again, general algorithms can be provided, but it should be expected that specific heuristic are eventually needed in order to obtain a better performance.

In present study, a very simple adaptation for such problem is proposed, referred as AVG_TS_SMOTE -the name stands on the idea of the generation of a new TS as the average of the two parents-. The selection of the parents is performed by random selection among the TS samples belonging to the minority class -as in SMOTE-. The generation of a new TS sample is performed as follows:

- For each feature, the average of the corresponding TS from the parents is computed.
- Each of the values of the class TS is calculated as the maximum of the values from the two parents.
- The length of the new TS sample, for every feature and for class TS, is bounded to the shortest of the two parents.

4 Experiments and Results

4.1 Experimental Setup

For this experimentation, a real world TS dataset obtained from the simulation of epileptic seizures is used; this dataset is publicly available at [3,19]. An epileptic seizure is a clinical manifestation that has its origin in abnormal electrical activity from groups of cortical neurons of variable size. Basically, there are two main types of epileptic seizures: generalized seizures and focal seizures. In both, there are subtypes with and without motor activity. In this study, we focus on the focal myoclonic seizure -repeated bursting movements of one limb, the upper and lower limbs of one body side or a combination of limb and facial movements.

The above referred TS dataset was gathered following a previously defined and very strict protocol, defining a set of activities, namely, the simulation of the epileptic convulsions and three activities: running, sawing and walking - either gesturing while walking slowly or normal walking at different paces. A wearable triaxial accelerometer sensor (3DACM) included in a bracelet placed on the affected wrist measured the participant movements.

The bracelets have wireless data sampling capabilities at a rate of 16 Hz, having the 3DACM a range of $2 \times g$. Up to 6 healthy participants, all of whom remained anonymous, successfully completed this experiment, each running 10 trials of each activity. The ages of the participants ranged from 22 to 47, with four participants of around 40 years old. One participant out of six was female, and the eldest was left-handed. An identification number was given to each Time Series (TS), including information fields on participant ID, the number of trials, the activity, etc.

The acceleration has been filtered and processed, becoming into a three variable TS dataset: the features are depicted in Table 1: the Signal-Magnitude Area (SMA), the Amount of Movement (AoM) and the Time between Peaks (TbP). The complete pre-processing have been described in [3].

This TS dataset, consisting on TS samples of three TS each -SMA, AoM and TbP- $\{\overline{TS}_s\}$, with the label for each activity $\{c_s\}$ and with the TS for each timestamp label $\{\overline{C}_s\}$, has been used in this experimentation.This original TS dataset is named ORIG, while the TS dataset after applying AVG_TS_SMOTE is named SMT.

To select the number of TS samples to be added to the dataset, the following criteria was used. In an imbalanced dataset, there exists $R = 3$ times more examples belonging to the MC class than to the mC class for the s data source. So, balancing the number of samples for both classes means injecting $(R1) \times |mC_s|$ new TS samples belonging to the mC.

The next experimentation focuses on analyzing the correlation between each feature and the class for the ORIG and the SMT datasets. The resulting TS dataset includes data from 6 participants, and for each participant up to 40 TS samples are included. Each TS sample includes, as stated in the introduction, three TS features -SMA, AoM and TbP-, the class label and the class TS.

Transformation	Calculation
$SMA_t(\boldsymbol{s})$	$\frac{1}{w}\sum_{i=1}^{w-1}(\sum_{c\in\{x,y,z\}}\|b_{c,t-1}\|)$
$AoM_t(\boldsymbol{s})$	$\sum_{i=0}^{i=w-1}\sum_{c\in\{x,y,z\}}\|max(b_{c,t-i}) - min(b_{c,t-i})\|$
$TbP_t(\boldsymbol{s})$	Computed with the following algorithm: 1.- Find the sequences with value higher than mean+K*std within the window ($K = 0.9$) 2.- Keep the rising points from each of these sequences 3.- Measure the mean time between them

Table 1. The transformations of the components of the acceleration, where $b_{c,i}$ stands for the body acceleration.

4.2 Correlation between each feature and the class

Two different measurements have been applied in this study in order to assess the relationship between the distribution of the ORIG and the SMT datasets, namely: the Pearson Correlation ($\rho_{X,Y}$, Eq. 1) coefficient and the Mutual Information ($MI(X,Y)$, Eq. 2); where cov is the covariance, σ_X is the standard deviation of X, $p(x)$ is the probability of the event x and $p(x,y)$ is the conditional probability of x given y.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

$$MI(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \tag{2}$$

The aggregated results for the six participants are shown in Table 2; however, only the boxplot for participant number 1 is included in Fig. 1. Table 3 shows the Wilcoxon test obtained results. The null hypothesis is that the data obtained for the $\rho(feature, class)$ or the $MI(feature, class)$ calculated for the ORIG dataset or calculated for the SMT dataset belongs to the same distribution. These results suggest nothing found against the truth of the null hypothesis.

Fig. 2 shows an example of how the new (synthetic) TS have been generated: the ability of the method to generate TS that highly resembles the original parents is remarkable. Nevertheless, in this study all the TS have been sampled using a very strict protocol, with a high control of the timing in each stage. Therefore, the TS do not differ too much from one to another; in other words, the generation of new synthetic TS was not as challenging as in other situations. However, this is an issue when facing the same problem of generating synthetic TS in different domains, or even in the same domain, in different contexts -such as in everyday life- where the TS are not so similar, even totally different. In these cases, different strategies for the generation of TS need to be addressed.

On the other hand, these new strategies should also consider the time and the complexity in generating a new TS sample. The main reason for that is that

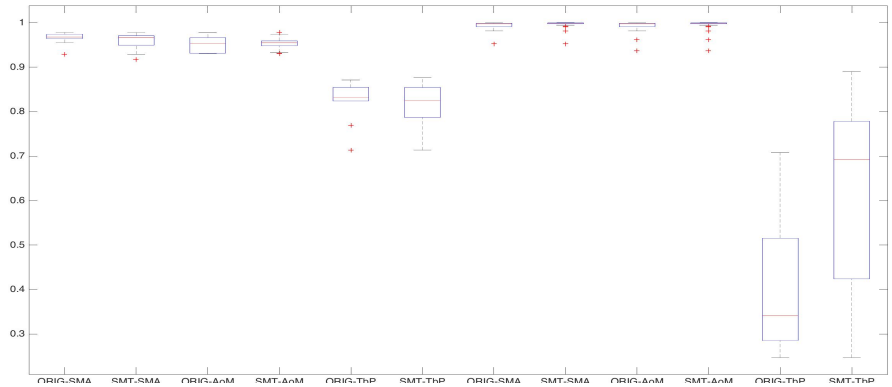

Fig. 1. Correlations between the class and the features for participant 1. From left to right, first for ORIG and then for SMT datasets, the results for ORIG-SMA, SMT-SMA, ORIG-AoM, SMT-AoM, ORIG-TbP and SMT-AoM.

a TS may be a short sequence, but it can be long enough. Any greedy strategy may work in the first case, but the longer the TS the worse the performance would be. Therefore, analysing energy efficiency concepts on the design of the TS generation strategy would eventually introduce more robust an scalable solutions.

5 Conclusions

In this study, the problem of balancing TS datasets is faced. Despite the effort in the development of balancing algorithms, the problem has been barely studied for the case of sets of TS. In this research, a simple yet efficient extension for the well-known SMOTE algorithm is detailed. The AVG_TS_SMOTE extension is based on three ideas: i) For each feature, the average of the corresponding TS from the parents is computed; ii) The class TS is calculated with the maximum of the values from the two parents; and iii) The length of the new TS sample, for every feature and for class TS, is bounded to the shortest of the two parents.

The experimentation shows the new balanced TS dataset is statistically similar to the original one when measured with the Pearson Correlation, while at the same time the new synthetic TS perform realistically and homogeneously. However, more research is needed to obtain a valid TS generation method that maintains the TS dataset statistical information. Nevertheless, two main concerns arose after: on the one hand, the need of good TS generation strategies on uncontrolled contexts; on the other hand, the requirement of considering energy efficiency issues in the design of such strategies. These two concerns are to be addressed in future work.

	Pearson Correlation					
	SMA		AoM		TbP	
Participant	**ORIG**	**SMT**	**ORIG**	**SMT**	**ORIG**	**SMT**
1	0.97/0.01	0.96/0.02	0.95/0.02	0.95/0.01	0.82/0.05	0.82/0.05
2	0.97/0.01	0.96/0.02	0.94/0.03	0.95/0.02	0.83/0.03	0.83/0.03
3	0.96/0.02	0.96/0.02	0.92/0.04	0.93/0.03	0.73/0.06	0.76/0.06
4	0.98/0.00	0.98/0.01	0.96/0.01	0.97/0.01	0.81/0.07	0.83/0.05
5	0.96/0.01	0.95/0.02	0.94/0.01	0.94/0.02	0.78/0.10	0.80/0.08
6	0.97/0.01	0.96/0.01	0.97/0.02	0.98/0.02	0.82/0.03	0.82/0.03
	Mutual Information					
	SMA		AoM		TbP	
Participant	**ORIG**	**SMT**	**ORIG**	**SMT**	**ORIG**	**SMT**
1	0.99/0.02	0.99/0.01	0.99/0.02	0.99/0.01	0.40/0.13	0.62/0.20
2	0.98/0.01	0.99/0.01	0.92/0.04	0.95/0.04	0.20/0.09	0.63/0.23
3	0.98/0.04	0.99/0.02	0.94/0.06	0.97/0.04	0.23/0.07	0.53/0.24
4	0.99/0.00	0.99/0.00	0.96/0.04	0.98/0.03	0.35/0.18	0.64/0.23
5	0.97/0.02	0.98/0.02	0.95/0.05	0.97/0.04	0.12/0.05	0.48/0.31
6	0.99/0.01	0.99/0.01	0.97/0.05	0.98/0.04	0.66/0.07	0.78/0.10

Table 2. Correlation measurements for the ORIG and SMT datasets, for each participant. Each cell includes the mean and the standard deviation statistics.

6 Acknowledgment

This research has been funded by the Spanish Ministry of Science and Innovation, under project MINECO-TIN2014-56967-R.

References

1. Beniczky, S., Polster, T., Kjaer, T., Hjalgrim, H.: Detection of generalized tonic-clonic seizures by a wireless wrist accelerometer: a prospective, multicenter study. Epilepsia **4**(54) (2013) e58–61
2. Villar, J.R., González, S., Sedano, J., Chira, C., Trejo-Gabriel-Galán, J.M.: Improving human activity recognition and its application in early stroke diagnosis. International Journal of Neural Systems **25**(4) (2015) 1450036–1450055
3. Villar, J.R., Vergara, P., Menéndez, M., de la Cal, E., González, V.M., Sedano, J.: Generalized models for the classification of abnormal movements in daily life and its applicability to epilepsy convulsion recognition. accepted for publication, International Journal of Neural Systems (2016)
4. Villar, J.R., Menéndez, M., de la Cal, E., González, V.M., Sedano, J.: Identification of abnormal movements with 3d accelerometer sensors for its application to seizure recognition. accepted for publication, International Journal of Applied Logic (2016)
5. López, V., Fernández, A., del Jesus, M., Herrera, F.: A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline data-sets. Knowledge-Based Systems **38** (2013) 85–104
6. Galar, M., Fernández, A., Barrenechea, E., Herrera, F.: Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern Recognition **46**(12) (2013) 3460–3471

	Pearson Correlation					
	SMA		AoM		TbP	
	ORIG	SMT	ORIG	SMT	ORIG	SMT
ORIG	1.0000	0.0000	1.0000	0.0068	1.0000	0.4047
SMT	0.0000	1.0000	0.0068	1.0000	0.4047	1.0000
	Mutual Information					
	SMA		AoM		TbP	
	ORIG	SMT	ORIG	SMT	ORIG	SMT
ORIG	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
SMT	1.0000	0.0000	0.0000	1.0000	0.0000	1.0000

Table 3. Wilcoxon signed-rank test results for participant 1 at a significance level of 0.05.

7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research (2002) 321–357
8. Batista, G., Prati, R., Monard, M.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations (2004) 20–29
9. He, H., Bai, Y., Garcia, E., Li, S., et al.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, IEEE (2008) 1322–1328
10. Tang, S., Chen, S.: The generation mechanism of synthetic minority class examples. In: Proceedings of 5th International Conference on Information Technology and Applications in Biomedicine (ITAB 2008). (2008) 444–447
11. Stefanowski, J., Wilk, S.: Selective pre-processing of imbalanced data for improving classification performance. In: Proceedings of the 10th International Conference in Data Warehousing and Knowledge Discovery (DaWaK2008). Volume LNCS 5182., Springer (2008) 283–292
12. Fu, T.c.: A review on time series data mining. Engineering Applications of Artificial Intelligence **24**(1) (2011) 164–181
13. Mishra, S., Saravanan, C., Dwivedi, V., Pathak, K.: Discovering flood rising pattern in hydrological time series data mining during the pre monsoon period. Indian Journal of Marine Sciences **44**(3) (2015) 3
14. Montgomery, D.C., Jennings, C.L., Kulahci, M.: Introduction to time series analysis and forecasting. John Wiley & Sons (2015)
15. Moses, D., et al.: A survey of data mining algorithms used in cardiovascular disease diagnosis from multi-lead ecg data. Kuwait Journal of Science **42**(2) (2015)
16. Köknar-Tezel, S., Latecki, L.J.: Improving svm classification on imbalanced time series data sets with ghost points. Knowledge and information systems **28**(1) (2011) 1–23
17. Agrawal, A., Viktor, H.L., Paquet, E.: Scut: Multi-class imbalanced data classification using smote and cluster-based undersampling. In: Proceedings of 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K). (2015)
18. Phan, S., Famili, F., Tang, Z., Pan, Y., Liu, Z., Ouyang, J., Lenferink, A., Oconnor, M.M.C.: A novel pattern based clustering methodology for time-series microarray data. International Journal of Computer Mathematics **84**(5) (2007) 585–597

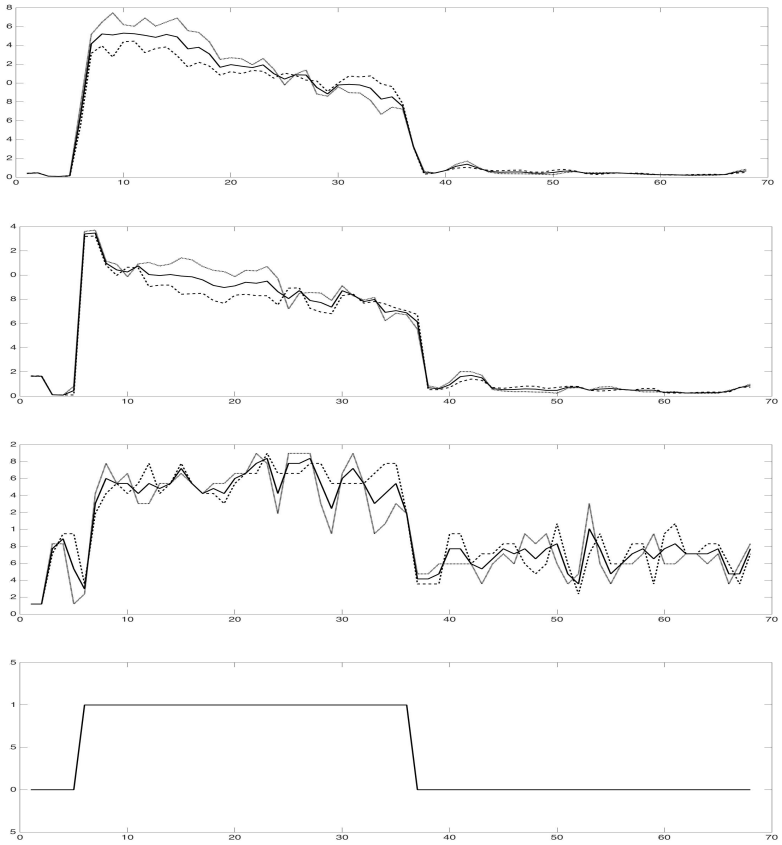

Fig. 2. Example of an AVG_TS_SMOTE Synthetic TS. From top to bottom, the performance on the SMA, AoM, TbP and class TS. For each figure, the two parents TS are plot with the dashed and the dotted lines, while the new synthetic TS is the solid line.

19. Villar, J.R.: Researcher's web page. `http://www.di.uniovi.es/~villar` (2017)