Comparison of Machine Learning Methods to Detect Anomalies in the Activity of Dairy Cows

Nicolas Wagner[1,2,3], Violaine Antoine[1], Jonas Koko[1], Marie-Madeleine Mialon[2], Romain Lardy[2], and Isabelle Veissier[2]

[1] UCA, LIMOS, UMR 6158, CNRS, Clermont-Ferrand, France
[2] UCA, INRAE, UMR Herbivores, F-63122 Saint-Genès-Champanelle, France
[3] nicolas.wagner@uca.fr

Abstract. Farmers need to detect any anomaly in animals as soon as possible for production efficiency (e.g. detection of estrus) and animal welfare (e.g. detection of diseases). The number of animals per farm is however increasing, making it difficult to detect anomalies. To help solving this problem, we undertook a study on dairy cows, in which their activity was captured by an indoor tracking system and considered as time series. The state of cows (diseases, estrus, no problem) was manually labelled by animal caretakers or by a sensor for ruminal pH (acidosis). In the present study, we propose a new Fourier based method (FBAT) to detect anomalies in time series. We compare FBAT with the best machine learning methods for time series classification in the current literature (BOSS, Hive-Cote, DTW, FCN and ResNet). It follows that BOSS, FBAT and deep learning methods yield the best performance but with different characteristics.

Keywords: Machine learning · Deep learning · Time series classification · Detection of anomalies · Precision livestock farming

1 Introduction

Precision livestock farming is based on the use of smart technologies (mainly sensors) to monitor closely the animals or their environment. The aim is to optimize the production and reduce farmers work load. The increase in computers storage capacity and in the precision of sensors makes possible to record a high quantity of data which requires automatic processing to be used by farmers. Machine learning tools is beginning to be employed to extract relevant information from these massive data. For example, machine learning has been used to determine grass growth from satellite and weather data [10] or to predict the quantity of manure to be spread on pastures or crops as fertilizer [11].

Farmers need to detect any anomaly in animals as soon as possible both for milk production efficiency and animal welfare. Such a detection seems possible through the analysis of the animals' activities. For instance, [15] found that dairy cows' activity varies according to a circadian cycle which significantly changes if the cow is about to be sick or in estrus.

Furthermore, time series classification (TSC) or anomaly detection are among the most challenging problem in machine learning [12], [18], [6] and are present in many fields of science like sensor-based human activity recognition [16], credit card fraud detection [1], electroencephalogram and electrocardiogram analysis [3], geo-distributed networks [5], etc. TSC differs from classical machine learning problems since it deals with data listed in time order. Some algorithms were developed for TSC like Dynamic Time Warping (DTW) [4], Bag of SFA Symbols (BOSS) [14] or Hive-Cote [8]. Most recently, deep learning neural networks as FCN and ResNet were also used and found to outperform the other algorithms [17], [6].

In this paper, we employed algorithms of time series classification (BOSS, DTW, Hive-Cote FCN and ResNet) considered as the best ones. In addition, because the activity of cows follows a circadian cycle, we proposed and tested a new method based on Fourier transformations (Fourier Based Approximation with Thresholding or FBAT).

The first section describes the most popular TSC classifier. The section 2 details our new FBAT method. Section 3 first describes the data set, i.e. time series of activities of dairy cows, then explains the experimental protocol and finally presents the results. Perspectives of the work are given in a conclusion.

2 Time Series Classifier

This section presents the current best algorithms for TSC [2], [17], [6] used as baseline to compare the FBAT method.

2.1 Dynamic Time Warping

DTW [4] is a method that measures the similarity between two time series. It is often used as a distance with the one Nearest Neighbor algorithm (1-NN). Although combining 1-NN and DTW gives good results in practice, however, DTW is not a distance function. Indeed, it does not respect all mathematical properties of a distance especially the triangle inequality [13].

The difference between DTW and standard distance measures is the following: standard distances assume that the i^{th} of a series is aligned with the i^{th} point of an other series while DTW is designed to minimize the effects of shifting and distortion in time series. The DTW method have many advantages. It is easy to employ, it can be used with many algorithms like k-NN and when combined with 1-NN it is one of the best algorithms for time series classification [2]. This makes it an interesting baseline. Its quadratic time complexity is a disadvantage but it remains faster than other algorithms like Hive-Cote.

2.2 Hive-Cote

Hive-Cote [8] is an improved version of the algorithm Cote (or Flat-Cote). Flat-Cote consists in using 35 classifiers of time series classification. Each classifier

produces a result and the final decision is based on a vote of all classifiers. The vote is weighted by the training accuracy of each classifier. One problem with Flat-Cote is the flat architecture. It means that all classifiers vote independently. However, some algorithms pertain to the same category and probably give similar results. To solve this problem, Hive-Cote gathers the algorithms into groups called modules. Each module computes the probability for each class to be the solution. This probability is computed with the weighted results of the algorithms for each module. Then, the final solution is the class that has the highest probability across all module's outputs.

Hive-cote is composed of five modules: elastic ensemble, shapelet transform ensemble, BOSS, time series forest and random interval features.

2.3 Fully Convolutional Networks

Fully Convolutional Networks (FCNs) [9] are similar to Convolutional Neural Networks (CNNs) excepted that they contain local pooling layer so as to keep the same dimensionality input through the convolutional layers. In addition, a standard CNN generally ends by a Fully Connected (FC) layer that is replaced by a Global Average Pooling (GAP) in the FCNs. The architecture used in this paper was proposed by [17]. It consists of three parts that are composed of a convolution layer, a Batch Normalization (BN) layer and a ReLu activation layer. These three parts are followed by a GAP layer and a classical softmax layer. The three convolution blocks contain 128, 256 and 128 filters with a filter length of respectively 8, 5 and 3. The stride is set to one with a zero padding that enables to preserve the same length of time series across the network. This architecture has the advantage to remain stable according to the length of the time series (excepted for the last softmax layer). This allows us to use exactly the same network used in [17] and [6]. FCN is the best deep learning algorithm on the 44 data sets analyzed in [17].

2.4 ResNet

A Residual Network [7] is close to CNN. The difference lies in shortcuts added from the input of convolution blocks to their output. These shortcuts inject the information that may be lost by the convolutional block. The ResNet proposed by [17] and [6] is composed of three convolutional blocks. All blocks are composed of three convolutional layers with respectively a filter's length set to 8, 5 and 3. Each convolutional layer is followed by a BN and Relu layer. The convolutional layers of the first block are composed of 64 filters and the convolutional layers of the second and last block are composed of 128 filters. The three blocks are followed by a global pooling and a softmax layer. This architecture has the same advantage as FCN: it does not vary with the length of the time series. ResNet is the best deep learning algorithm on the 85 data sets tested in [6].

2.5 Bag of SFA Symbols

BOSS [14] is a method that combines the advantage of the Fourier transform and the bag of words model. It allows to reduce noise and to handle variable lengths.

First, a sliding window of size w with a step of 1 is applied over each time series. The obtained time windows are converted into sequences (or words) of symbols of length l with an alphabet size of c using the Symbolic Fourier Approximation (SFA) algorithm. A time series is then represented by the sequences of each window. Finally, a histogram is built using sequences as modal class. The last step consists in using 1-NN algorithm with the BOSS distance function. Given two histograms B_1 and B_2, the formula of the BOSS distance function is:

$$dist(B_1, B_2) = \sum_{a \in B_1; B_1(a) > 0} [B_1(a) - B_2(a)]^2, \tag{1}$$

where a is a word and $B_i(a)$ the number of occurrences of a in the i^{th} histogram.

3 Fourier Based Approximation with Thresholding

We propose a Fourier Based Approximation with Thresholding (FBAT) method to classify time series by measuring the variations of the cyclic components. It is made to classify time series as normal or abnormal by assuming that an abnormal series includes a break on the cycle. Thus, if the variations of the cyclic components are high, the algorithm classifies the time series S_i as abnormal. The algorithm starts by extracting two sub-series A and B of size p and delayed of q from the input series with $p < |S_i|$ and $p + q < |S_i|$. A Fourier transform is applied on both sub-series to extract their harmonic decomposition. With these harmonics, a new model $m(t)$ is computed for each sub-series following the formula:

$$m(t) = \sum_{f=-z}^{z} |h_f| cos(2\pi f \frac{t}{p} + arg(h_f)),\ z = 0...\lceil \frac{p-1}{2} \rceil, \tag{2}$$

where h_f is the harmonic corresponding to the frequency f and z is the number of harmonics to keep in the model. Note that h_f is a complex number with $|h_f|$ its modulus and $arg(h_f)$ its argument. Moreover, the two sub-series A and B are delayed by q. As a consequence, it is necessary to synchronize the models of A and B by applying a temporal shift to the model of B. This shift is performed by adding a delay $-\frac{q}{p}2\pi$ in the formula of the model of B.

A L2-norm distance d_{L2} is then computed between the two models. This distance reflects the variation of the cyclic component of the input time series. A high distance means a high variation and vice versa.

To classify the input time series as normal or abnormal, the algorithm needs to compute a threshold τ for the distance. If $d_{L2} > \tau$, the time series is classified as abnormal. To compute this threshold, all distances d_{L2} are computed for

each time series that belongs to the training set. Then, s samples are computed between the minimum and the maximum obtained distances. The accuracy of the training set is computed for each sample and the sample that yields the best accuracy is chosen to be the threshold.

4 Experiments and Results

4.1 Data set

The data were collected on 28 Holstein cows during a two month experimentation in which a subacute ruminal acidosis, a metabolic disease common in ruminants, was induced.

Data construction The raw data consist of the record of the location of each cow every second with an indoor tracking system (CowView, GEA Farm Technologies, Bönen, Germany). Three activities were identified: *eating* if the cow was located next to the trough, *resting* if it was in a cubicle (resting place) and *in alleys* if the cow was in an alley. These activities were aggregated in a new variable called level of activity. The procedure is described in [15]. We thus obtained for our study time series consisting in the evolution over time of the level of activity of each cow estimated per hour. All anomalies, such as acidosis, oestrus, etc were noted. The acidosis was detected by a sensor that measured the pH in the rumen.

A set of 28 time series, corresponding to the 28 cows for two months was available. To build the data set, a sliding window of 36 hours was applied on each cow to extract sub-series (see justification in next section). The obtained data set was divided into two parts: one for the training and one for test. Half of series labelled as *abnormal* were used for the training set and the other half for the test set, except for the series related to acidosis that were all placed in the test set. Indeed, this specific anomaly was induced by experimenters. The idea is to avoid the perturbation of a classifier during the learning phase with an unnatural anomaly. Finally, to balance the training data set, a reduction of the *normal* series was performed by randomly selecting few ones. The same number of normal series were randomly chosen for the test data set. The training data set is composed of 1088 *normal* series and 972 *abnormal* series. The test set is composed of 1408 *normal* series and 4212 *abnormal* series (including 3180 series with acidosis).

Data properties The first property of this data set is that each cow has its own natural daily rhythm based on a circadian cycle with a low activity during the night and a higher activity during the day [15]. This change of activity between nights and days can be modified if the cow is sick or under stress. We decided to work with series of 36 hours in order to observe a cycle of more than one day and to be able to detect anomalies with precision (e.g. a normal cycle of 24 hours followed by 12 hours abnormal).

Figure 1 illustrates the activity of two cows during two normal days. Both cows are more active during the day than at night, in line with the fact that the rhythm is circadian. This figure also illustrates that the rhythm of normal activity can differ between cows and days. Figure 2 illustrates the level of activity of a cow under lame during three days. This shows how the activity of a cow can be modified by an anomaly. Given that there exists normal variations between days (as illustrated by the Cow b, Fig. 1), the difficulty lies in discriminating between changes due to an anomaly and those due to spontaneous variations.

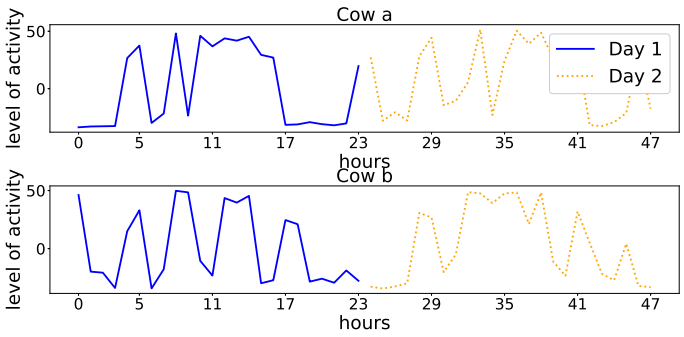
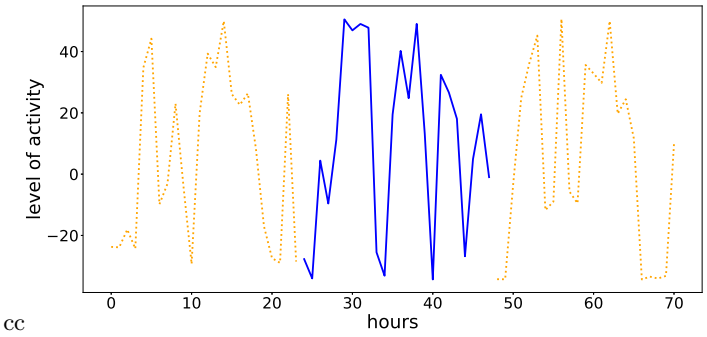


Fig. 1. 48 hours of normal activity for two different cows.



cc

Fig. 2. Level of activity of a cow during three consecutive days: the solid line represents the day detected as lame.

4.2 Experiments

We compare the algorithms described in section 2 with the FBAT method presented section 3. The code of the five methods comes from the github repository

of the original authors [2], [6]. The code of the FBAT method is available on the github repository https://github.com/nicolas-wagner/FBAT.

For BOSS we use the same parameters as in [2]: the word length $l = [8, 10, 12, 14, 16]$, the alphabet size $c = 4$ and the window size, $w = [10, 12, ..., 36]$.

For FBAT we set the time window p to 24, the delay q to 12 and the number of samples s to 10000. We test all possibilities of the number of harmonics z, i.e. from 0 to 12 harmonics.

As in [6], the deep learning methods were run 10 times to train them with 10 different initializations of parameters. The results presented in this paper are the average over these 10 runs.

We use the same train and test data set for all methods tested. The train data set is composed of 1088 time series labelled as *normal* (negative) and 972 time series labeled as *abnormal* (positive). The test set is composed of 1408 *normal* time series and 4212 *abnormal*.

We define normal label as the negative class and abnormal label as the positive class. For each classifier it is possible to count the number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). We then calculated the overall accuracy as well as the precision and the recall for positive and the negative classes as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

$$precision_{-} = \frac{TN}{TN + FN} \quad \text{and} \quad recall_{-} = \frac{TN}{TN + FP}, \tag{4}$$

$$precision_{+} = \frac{TP}{TP + FP} \quad \text{and} \quad recall_{+} = \frac{TP}{TP + FN}. \tag{5}$$

The accuracy is used as a single value to measure and compare the performance of the methods. The precision and the recall measured for each class help to understand the behavior of each classifier in detail. A high $recall_{-}$ (resp. $precision_{-}$) means that the majority of time series labelled as normal (resp. classified as normal) are classified as normal (resp. labelled as normal) and inversely for a high $recall_{+}$ (resp. $precision_{+}$).

The CPU time (in hour) is also retrieved from the experiments (training + test time). For the deep learning methods, a GPU mode is available so, FCN and ResNet were run on CPU and GPU. The first machine were composed of CPUs Intel Xeon 2.4GHz with 80 cores and 1 TB of RAM. The second were composed of CPUs Intel Xeon 2.4GHz with 10 cores and 62.5 GB of RAM with a GPU NVIDIA Quadro P5000 (16 GB of GDDR5 memory and 2560 cores). All algorithms were run in a sequential mode using only one core.

4.3 Results

The performances of all methods are summarized Tab. 1.

Table 1. Results of all classifiers

		DTW	Hive-Cote	BOSS	FBAT	FCN	ResNet
accuracy		0.54	0.63	**0.72**	0.60	0.66	0.67
$precision_-$		0.31	0.38	**0.43**	0.38	0.40	0.40
$recall_-$		0.68	0.73	0.36	**0.90**	0.73	0.71
$precision_+$		0.82	0.87	0.80	**0.94**	0.88	0.87
$recall_+$		0.49	0.60	**0.84**	0.50	0.64	0.65
time (h)	CPU	1h10	28h	0h38	**0h06**	19h28	16h36
	GPU	-	-	-	-	1h16	2h13

All of these methods are intended to be used by livestock farmers with a personal computer. Consequently, the CPU time is an important characteristic to take into account and we notice that Hive-Cote has a too large CPU time (28 hours) to be used in real conditions.

DTW obtains the worst performance in terms of accuracy (0.54) and its CPU time is rather high. DTW is faster than Hive-Cote but slower than BOSS and FBAT and similar to the GPU time of the deep learning models. Therefore, we estimate that DTW does not obtain enough satisfying results to be kept as a solution for this problem.

The performances of the neural networks are similar, excepted in terms of GPU time where FCN is almost two times faster than ResNet. They are a compromise between BOSS and FBAT in terms of $recall_+$ and $recall_-$. However, FCN and ResNet are considered as expensive solutions for livestock farmers since they need a GPU to be used in a real application.

BOSS produces the best accuracy results and obtains a low recall of the negative class. This means that among all time series labelled as *normal*, most of them are incorrectly classified as *abnormal*. If a farmer decides to use BOSS as a tool for detecting anomalies in dairy cows, he/she will then receive a high number of false alerts. These wrong detections may overshadow the correct ones and the method can become worthless. On the opposite, FBAT has the higher recall for the negative class. This would lead to a low number of false alerts for the farmer. The recall of abnormal days ($recall_+$) metric however decreases from 0.84 for BOSS to 0.50 for FBAT. As a matter of fact, a low $recall_+$ score may be due to the data set construction. Indeed, thanks to previous observations [15], we chose to label all time series included between two days before and one day after an anomaly as abnormal because the behavior of an animal can be disturbed shortly before and after clinical symptoms are detected. But all anomalies may not last for four consecutive days and this can lower the $recall_+$. We checked if for each anomaly, at least one of the four days is detected as abnormal by FBAT. The FBAT method detected at least one day among the four consecutive ones labelled as abnormal in 83% of the lameness cases, 61% of the acidosis and 100% of the estrus. These results seem adequate for an on-farm use and a test by farmers is necessary to decide if the $recall_-$ is satisfactory.

Another advantage of the FBAT method in a farm application is the threshold. Indeed, we proposed a solution to automatically set the threshold between normal vs. abnormal time series. Nevertheless, the threshold can be adjusted. If a farmer thinks that the method does not detect enough anomalies, the threshold can be decreased. On the opposite, if the method detects too much false positive series, the threshold can be increased. Moreover, a threshold can be defined for each cow. If a cow is particularly insensitive to anomalies, its threshold can be decreased and inversely for a cow with a higher sensitivity.

The last advantage of FBAT is its CPU time, which is the best one of our experiments with 6 minutes. We also tested FBAT on personal computers (instead of big and expensive servers) and the CPU time didn't exceed 30 minutes.

5 Conclusion

The early detection of anomalies is very important for a farmer. Thanks to tools developed for precision livestock farming, it is possible to collect data in real time that can be analyzed by machine learning methods. In this study, we proposed a method based on Fourier transforms (FBAT) that we tested with the best algorithms and deep learning models available in the current literature for time series classification (BOSS, Hive-Cote, DTW, FCN and ResNet). The results showed that FBAT and BOSS are the two best solutions to solve the problem of anomaly detection in dairy cow activity. BOSS gives the best recall in the negative class whereas FBAT gives the best recall in positive class. They both obtain the best CPU time and they are both easy to implement. FBAT has the advantage to employ a threshold that can be adjusted to each cow. Testing these methods in real conditions, that is by farmers themselves, should help to choose the best method for the purpose. As other perspective, we propose to consider the labels as fuzzy. Indeed, the anomalies are detected when clinical signs are well visible, but it is reasonable to assume that anomalies gradually appear and disappear. We expect to increase the performances of the classifiers by better defining the labels. Finally, we plan to study the robustness of the algorithms to noisy labels. Indeed, label noise often occurs when humans are involved. In our application, caretakers are detecting and labeling anomalies but they easily have imperfect evidence.

Acknowledgment

This collaborative work was made possible thanks to the French Government IDEX-ISITE initiative 16-IDEX-0001 (CAP 20-25). The PhD grant for N. Wagner was provided by INRA and Université Clermont Auvergne. We thank the HERBIPOLE staff, B. Meunier, Y. Gaudron and M. Silberberg for data.

References

1. Adewumi, A.O., Akinyelu, A.A.: A survey of machine-learning and nature-inspired based credit card fraud detection techniques. International Journal of System Assurance Engineering and Management **8**(2), 937–953 (2017)
2. Bagnall et al.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery **31**(3), 606–660 (2017)
3. Berkaya, S.K., Uysal, A.K., Gunal, E.S., Ergin, S., Gunal, S., Gulmezoglu, M.B.: A survey on ecg analysis. Biomedical Signal Processing and Control **43**, 216–235 (2018)
4. Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: KDD workshop. vol. 10, pp. 359–370. Seattle, WA (1994)
5. Corizzo, R., Ceci, M., Japkowicz, N.: Anomaly detection and repair for accurate predictions in geo-distributed big data. Big Data Research **16**, 18–35 (2019)
6. Fawaz et al.: Deep learning for time series classification: a review. Data Mining and Knowledge Discovery **33**(4), 917–963 (2019)
7. He et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Lines, J., Taylor, S., Bagnall, A.: Hive-cote: The hierarchical vote collective of transformation-based ensembles for time series classification. In: 2016 IEEE 16th international conference on data mining (ICDM). pp. 1041–1046. IEEE (2016)
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
10. Marwah, R., Cawkwell, F., Hennessy, D., Green, S.: Improved estimation of grassland biomass using machine learning and satellite data. In: 9th ECPLF 2019. pp. 174–179. Teagasc (2019)
11. Mollenhors, H., de Haan, M., Oenema, J., Hoving-Bolink, A., Veerkamp, R., Kamphuis, C.: Machine learning to realize phosphate equilibrium at field level in dairy farming. In: 9th ECPLF 2019. pp. 41–44. Teagasc (2019)
12. Munir, M., Siddiqui, S.A., Dengel, A., Ahmed, S.: Deepant: A deep learning approach for unsupervised anomaly detection in time series. IEEE Access **7**, 1991–2005 (2018)
13. Ruiz, E.V., Nolla, F.C., Segovia, H.R.: Is the dtw "distance" really a metric? an algorithm reducing the number of dtw comparisons in isolated word recognition. Speech Communication **4**(4), 333–344 (1985)
14. Schäfer, P.: The BOSS is concerned with time series classification in the presence of noise. Data Mining and Knowledge Discovery **29**(6), 1505–1530 (2015)
15. Veissier, I., Mialon, M.M., Sloth, K.H.: Early modification of the circadian organization of cow activity in relation to disease or estrus. Journal of dairy science **100**(5), 3969–3974 (2017)
16. Wang, J., Chen, Y., Hao, S., Peng, X., Hu, L.: Deep learning for sensor-based activity recognition: A survey. Pattern Recognition Letters **119**, 3–11 (2019)
17. Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: A strong baseline. In: 2017 International joint conference on neural networks (IJCNN). pp. 1578–1585. IEEE (2017)
18. Yang, Q., Wu, X.: 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making **5**(04), 597–604 (2006)