Fuzzy k-NN based classifiers for time series with soft labels

Nicolas Wagner[1,2,3][0000−0002−5480−0500], Violaine Antoine[1][0000−0002−0981−3505], Jonas Koko[1][0000−0002−0970−5002], and Romain Lardy[2][0000−0003−1338−8553]

[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. Time series are temporal ordered data available in many fields of science such as medicine, physics, astronomy, audio, etc. Various methods have been proposed to analyze time series. Amongst them, time series classification consists in predicting the class of a time series according to a set of already labeled data. However, the performance of a time series classification algorithm depends on quality of the known labels. In real applications, the time series are often labelled by an expert or by an imprecise process, leading to noisy labels. Several algorithms have been developed to handle uncertain labels in case of non-temporal data sets. As an example, the fuzzy k-NN introduce for labeled objects a degree of membership to belong to classes. In this paper, we combine two popular time series classification algorithms, Bag of SFA Symbols (BOSS) and the Dynamic Time Warping (DTW) with the fuzzy k-NN. Results show that our fuzzy time series classification algorithms outperform the non-soft algorithms especially when the level of noise is high.

Keywords: time series classification · BOSS · fuzzy k-NN · soft labels.

1 Introduction

Time series (TS) are data constrained with time order. Such data frequently appear in many fields such as economics, marketing, medicine, biology, physics... There exists a long-standing interest for time series analysis methods. Amongst the developed techniques, time series classification attract much attention since the need to accurately forecast and classify time series data spanned across a wide variety of application problems [2, 20, 9].

A majority of time series approaches consists in transforming time series and/or creating an alternative distance measure in order to finally employ a basic classifier. Thus, one of the most popular time series classifier is a *k-Nearest Neighbor* (k-NN) using a similarity measure called *Dynamic time warping* (DTW) [12] that allows non linear mapping. More recently, a *bag-of-words* model combined with the *Symbolic Fourier Approximation* (SFA) algorithm [19] has been developed in order to deal with extraneous and erroneous data [18]. The algorithm, referred to as Bag of SFA Symbols (BOSS), converts time series into histograms.

A distance is then proposed and applied to a k-NN classifier. The combinations of DTW and BOSS with a k-NN are simple and efficient approaches used as gold standards in the literature [1, 8].

The k-NN algorithm is a lazy classifier employing labeled data to predict the class label of a new data point. In time series, labels are specified for each timestamp and are obtained by an expert or by a combination of sensors. However, changing from one label to another can span multiple timestamps. For example in animal health monitoring, an animal is more likely to become sick gradually than suddenly. As a consequence, using soft labels instead of hard labels to consider the animal state seems more intuitive.

The use of soft labels in classification for non time series data sets has been studied and has shown robust prediction against label noise [7, 21]. Several extensions of the k-NN algorithm have been proposed [6, 10, 14]. Amongst them, the fuzzy k-NN [11], which is the most popular algorithm [5], handles labels with probabilities membership for each class. The fuzzy k-NN has been applied in many domains: in bioinformatics [22], image processing [13], fault detection [24], etc.

In this paper, we consider the most popular time series algorithms that is the k-NN classifier and we propose to replace by a fuzzy k-NN. The purpose is to tackle the problem of gradual labels in time series.

The rest of the work is organized as follows. Section 2 first recals the DTW and BOSS algorithms. Then, the fuzzy k-NN classifier as well as the combinations between BOSS/DTW and fuzzy k-NN are detailed. Section 3 presents a comparison between hard and soft labels through several data sets. Section 4 concludes the paper.

2 Time series classifiers for soft labels

2.1 Dynamic time warping (DTW)

Dynamic Time Warping [3] is one of the most famous similarity measurement between two times series. It takes into account the fact that two similar times series may have different lengths due to various speed. The DTW measure allows then a non-linear mapping, which implies a time distortion. It has been shown that DTW is giving better comparisons than a Euclidean distance metric. In addition, the combination of the elastic measure with the 1-NN algorithm is a gold standard that produces competitive results [1], although DTW is not a distance function. Indeed, DTW does not respect the property of triangle inequality but in practice, this property is often respected [17]. Despite DTW has a quadratic complexity, the use of this measure with a simple classifier remains faster than other algorithms like neural networks. Moreover, using lower bound technique can decrease the complexity of the measure to a linear complexity [16].

2.2 The Bag of SFA Symbols (BOSS)

The bag of SFA Symbols algorithm (BOSS) [18] is a bag of words method using Fourier transforms in order to reduce noise and to handle variable lengths. First,

a sliding window of size w is applied on each time series of a data set. Then, windows from the same time series are converted into a word sequences according to the Symbolic Fourier Approximation (SFA) algorithm [19]. Words are composed of l symbols with an alphabet size of c. The time series is then represented by an histogram that corresponds to the number of word occurrences for each word. Finally, the 1-NN classifier can be used with distance computed between histograms. Given two histograms B_1 and B_2, the measure called d_{BOSS} is:

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

where a is a word and $B_i(a)$ the number of occurrences of a in the i^{th} histogram. Note that the set of words are identical for B_1 and B_2, but the number of occurrences for some words can be equal to 0.

2.3 Fuzzy k-NN

Let $\mathcal{D} = (\mathcal{X}, y)$ be a data set composed of $n = |\mathcal{X}|$ instances and $y_i \in \mathcal{C}$ be a label assigned to each instance $x_i \in \mathcal{X}$ with $\mathcal{C}$ the set of all possible labels.

For conventional hard classification algorithms, it is possible to compute a characteristic function $f_c : \mathcal{X} \to \{0, 1\}$ with $c \in \mathcal{C}$:

$$f_c(x_i) = \begin{cases} 1, c = y_i, \\ 0, c \neq y_i. \end{cases} \tag{2}$$

Rather than hard labels, soft labels allow to express a degree of confidence on the class membership of an object. Most of the time, this uncertainty is represented given by probabilistic distribution. In that case, soft labels corresponds to fuzzy labels. Thereby, the concept of characteristic function is generalized to membership function $u_c : \mathcal{X} \to [0, 1]$ with $c \in \mathcal{C}$:

$$u_c(x_i) = \mathcal{P}(y_i = c), \tag{3}$$

such that

$$\sum_{c \in \mathcal{C}} u_c(x_i) = 1, \tag{4}$$

$$0 < \sum_{x \in \mathcal{X}} u_c(x) = 1 < n, \forall c \in \mathcal{C}. \tag{5}$$

There exists a wide range of k-NN variants using fuzzy labels in the literature [5]. The most famous and basic method, referred to as fuzzy k-NN [11], predicts the class membership of an object x_i using two steps. First, similarly to the hard k-NN algorithm, the k nearest neighbors $x_j \in \mathcal{K}$, $|\mathcal{K}| = k$ of x_i are retrieved. The second step differs from hard k-NN as it computes a membership degree for each class:

$$u_c(x_i) = \frac{\sum_{x_j \in \mathcal{K}} u_c(x_j) d(x_i, x_j)^{-2/(m-1)}}{\sum_{x_j \in \mathcal{K}} d(x_i, x_j)^{-2/(m-1)}}, \forall c \in \mathcal{C}, \quad (6)$$

with m a fixed coefficient controlling the fuzziness of the prediction, $d(x_i, x_j)$ the distance between instances x_i and x_j. Usually, $m = 2$ and the Euclidean distance is the most popular distance considered.

2.4 Fuzzy DTW & Fuzzy BOSS

In order to deal with time series and fuzzy labels, we propose two fuzzy classifiers called F-DTW and F-BOSS.

The F-DTW algorithm consists in using the fuzzy k-NN algorithm with DTW as distance function (see Fig. 1). It takes in entry a time series and computes the DTW distance with the labeled times series. Once the k closest time series found, the class membership is computed with equation (6).

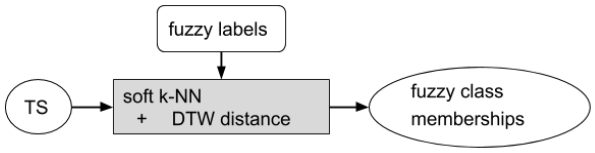

Fig. 1. F-DTW algorithm

The F-BOSS algorithm consists in first applying the BOSS algorithm in order to transform the time series into histograms. Then, the fuzzy k-NN is applied with BOSS distances. It generates fuzzy class memberships (see Fig. 2).

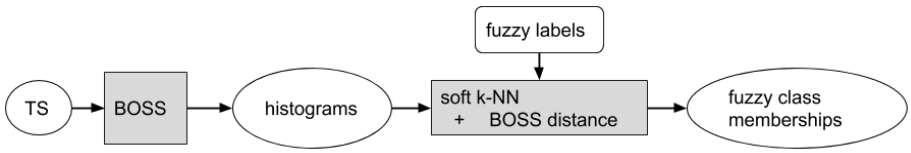

Fig. 2. F-BOSS algorithm

3 Experiments

3.1 Experimental protocol

We have selected four data sets from the University of California Riverside (UCR) archive [4]. Each data set have different characteristics detailed Table 1.

Table 1. Characteristics of data sets.

Data set name	Size train	Size test	Size series	Nb classes	Type
WormsTwoClass	181	77	900	2	MOTION
Lightning2	60	61	637	2	SENSOR
ProximalPhalanxTW	400	205	80	6	IMAGE
Yoga	300	3000	426	2	IMAGE

The hard labels are known for each data set. Thus, we generate fuzzy labels as described in [15]. First noise is introduced in the label set in order to represent uncertain knowledge: for each instance x_i, a probability p_i to alter label y_i is randomly generated according to a beta distribution with a variance σ set to $\sigma = 0.04$ and the expectation μ set to $\mu = [0.1, 0.2, ..., 0.7]$. In order to decide if the label of x_i is modified, another random number p'_i is generated according to an uniform distribution. If $p_i > p'_i$, a new label $y'_i \in \mathcal{C}$ such that $y'_i \neq y_i$ is randomly assigned to x_i. Second, fuzzy labels are deduced using p_i. Let $\Pi_c : \mathcal{X} \rightarrow [0, 1]$ be a possibilistic function computed for each instance x_i and each class c:

$$\Pi_c(x_i) = \begin{cases} 1, & c = y'_i, \\ p_i, & c \neq y'_i. \end{cases} \tag{7}$$

The possibilistic distribution allows to go from total certainty when $p_i = 0$ to total uncertainty when $p_i = 1$. Since our algorithms employ fuzzy labels, possibilities Π_i are converted into probabilities u_c by normalizing equation (7) with the sum of all possibilities:

$$u_c(x_i) = \frac{\Pi_c(x_i)}{\sum_{c \in \mathcal{C}} \Pi_c(x_i)}. \tag{8}$$

We propose to test and compare three strategies dealing with noisy labels. The two first ones are dedicated to classifiers taking in entry hard labels.

The first strategy, called strategy 1, considers that noise in labels is unknown. As a result soft labels are ignored and for each instance x_i, label y_i^* is chosen using the maximum probability membership rule, i.e. $\max(u_c(x_i))$.

The second strategy, called strategy 2, consists in discarding the most uncertain labels and transforming soft labels into hard labels. For each instance x_i the normalized entropy H_i is computed as follows:

$$H_i = \frac{1}{\log_2(|\mathcal{C}|)}(-\sum_{k \in \mathcal{C}} u_k(x_i) \log_2(u_k(x_i))). \tag{9}$$

Note that $H_i \in [0, 1]$ and $H_i = 0$ corresponds to a state of total certainty whereas $H_i = 1$ corresponds to an uniform distribution. If $H_i > \theta$ we consider the soft label of x_i as too uncertain and x_i is discarded from the fuzzy data set. In the experiments, we set the threshold θ to 0.95.

Finally, the third strategy, called strategy 3, keeps the whole fuzzy labels and apply a classifier able to handle such labels.

In order to compare strategies and since strategies 1 and 2 give hard labels whereas strategy 3 generates fuzzy labels, we convert fuzzy labels using the maximum membership rule, i.e. $\max(u_c(x_i))$, $\forall c \in \mathcal{C}$.

The best parameters of F-BOSS are found by a leave-one-out cross-validation on the training set. The values of the parameters are fixed as in [1]:

- window length $w = [10, ..., q]$, with $q = |x_i|$, the size of the series and $|w| = \min(200, \sqrt{q})$,
- alphabet size $\alpha = 4$,
- word length $l = [8, 10, 12, 14, 16]$.

Classifiers tested are soft k-NN, F-BOSS and F-DTW. For strategies 1 and 2, they correspond to k-NN, BOSS with k-NN and DTW with k-NN. For each classifier, different numbers of neighbors $k = [1, 2, ..., 10]$ and different values of μ, $\mu = [0, 0.1, 0.2, ..., 0.7]$ are analyzed. Note that $\mu = 0$ corresponds to the original data set without fuzzy processing. To compare the differents classifiers and strategies, we choose to present the percentage of good classification, referred to as accuracy.

3.2 Influence of the number of neighbors in k-NN

Usually with DTW or BOSS with hard labels, the number of neighbors is set to 1. This experiment studies the influence of the parameter k when soft labels are used. Thus, we set $\mu = 0.3$ in order to represents a moderate level of noise that can exist in real applications and apply strategie 3 on all data sets. Figure 3 illustrates the result on the WormsTwoClass data set, i.e. the variation of the accuracy for the three classifiers according to k.

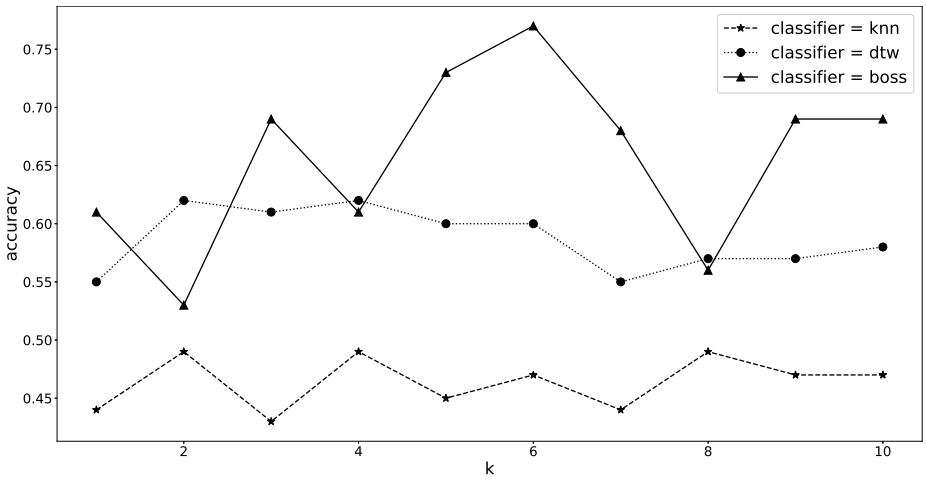

Fig. 3. Accuracy according to k for WormsTwoClass data set: $\mu = 0.3$ and strategy 3

First, for all values of k the performance of the soft k-NN classifier is under the others. Such result has also been identified in other data set. We also observe on Figure 3 that the F-BOSS algorithm is often better than F-DTW. However the pattern of the F-BOSS curve is serrated that makes difficult the establishment of guidelines for the choice of k. In addition, the best k depends on the algorithm and the data set. This is why, for the rest of the experiments section, we chose to set k to the median value $k = 5$.

3.3 Strategies and algorithms comparisons

Table 2 presents the results of all classifiers and all strategies on the four data sets for $k = 5$ and $\mu = 0.3$. The k-NN classifier has always the worst performances. This result is expected since DTW and BOSS algorithms are specially developed for time series problems. The best algorithm between F-DTW and F-BOSS depends on the data set: F-DTW is slightly the best one for Lightning2 and Yoga, and F-BOSS is the best one for WormsTwoClass. Note that for ProximalPhalanxTW, F-DTW is the best with strategy 2 and F-BOSS is the best for the third strategy. Strategy 1 (i.e. hard labels) is most of the time worst than the two other strategies. This can be explained by the fact that the first strategy does not take the noise into account. For all best classifiers of all data sets, the third strategy is the best strategy even though for ProximalPhalanxTW, strategy 2 competes with strategy 3. The third strategy (i.e. soft) is therefore most of the time better than the second (i.e. discard) one. However, the best algorithm between F-BOSS and F-DTW depends on the data sets.

Table 2. Accuracy for all data sets with $\mu = 0.3$ and $k = 5$.

	ProximalPhal.			Lightning2			WormsTwoC.			Yoga		
strategy	1	2	3	1	2	3	1	2	3	1	2	3
soft k-NN	0.32	0.38	0.38	0.43	0.59	0.56	0.44	0.48	0.45	0.64	0.68	0.68
F-DTW	0.33	**0.41**	0.39	0.67	0.67	**0.69**	0.56	0.58	0.6	0.68	0.72	**0.73**
F-BOSS	0.36	0.4	**0.41**	0.56	0.66	0.56	0.7	0.68	**0.73**	0.67	0.71	0.7

3.4 Noise impact on F-BOSS and F-DTW

To observe the impact of the μ parameter, Fig. 4 and Fig. 5 illustrate respectively the accuracy variations for the WormsTwoClass and Lightning2 data sets according to the value of μ. The k-NN classifier and the first strategy are not represented because their performances are not satisfying (see Section 3.3). The figures also include the value $\mu = 0$ that corresponds to the original data without fuzzy processing. Results are not presented for the Yoga and ProximalPhalanxTW data sets because the accuracy differences between the strategies and the classifiers are not significant, especially when $\mu < 0.3$.

For WormsTwoClass, the F-BOSS algorithm is better than F-DTW and inversely for Lightning2 data set. For the both data sets, with a low or moderate level of noise ($\mu < 0.3$), the third strategy is better than the second one. Higher levels of noise lead to better results with strategy 2. This can be explained as follows: strategy 2 is less disturbed by the important number of miss-classified instances since it removes them. On the opposite, with a moderate level of noise, the soft algorithms are more accurate because they keep informative labels.

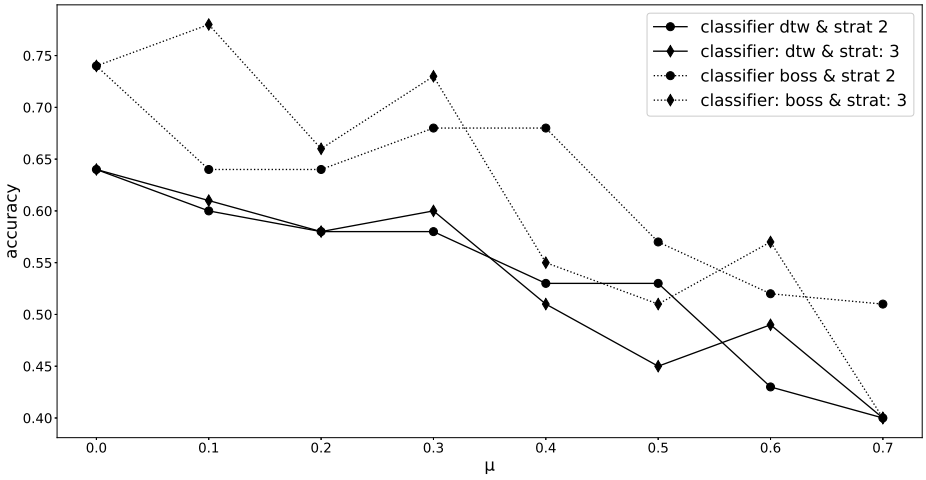

Fig. 4. Accuracy according to μ for WormsTwoClass data set

Predicting soft labels instead of hard labels brings to the expert an extra information that can be analyzed. We propose to consider as uncertain all predicted fuzzy labels having a probability less than a threshold t. Figure 6 present the accuracy and the number of elements discarded varying with this threshold t for the WormsTwoClass data set. As it can be observed, the higher is t, the better is the accuracy and the more the number of predicted instances are discarded. Thus t is a tradeoff between good results and a sufficient number of predicted instances.

4 Conclusion

This paper considers the classification problem of time series having fuzzy labels, i.e. labels with probabilities to belong to classes. We proposed two methods, F-BOSS and F-DTW, that are a combination of a fuzzy classifier (k-NN) and methods dedicated to times series (BOSS and DTW). The new algorithms are tested on four data sets coming from the UCR archives. With F-BOSS and F-DTW, integrating the information on uncertainty about the class memberships

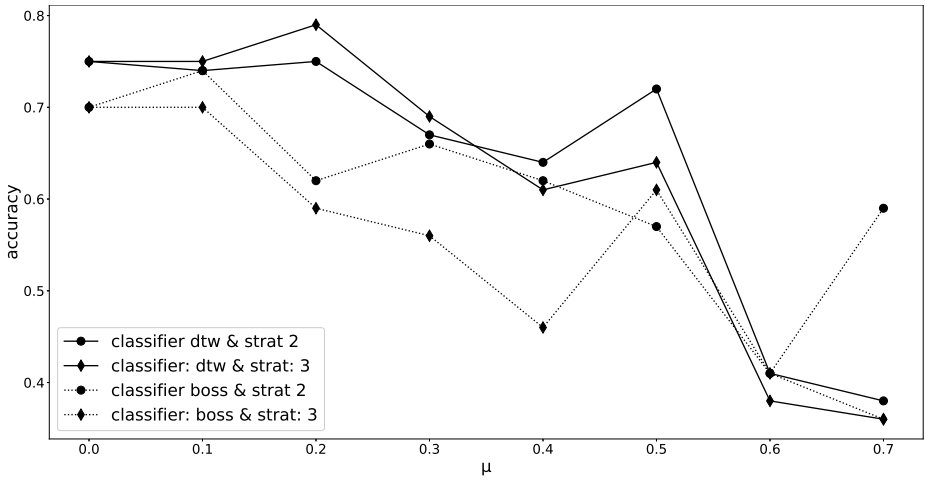

Fig. 5. Accuracy according to μ for Lightning2 data set: $k = 5$

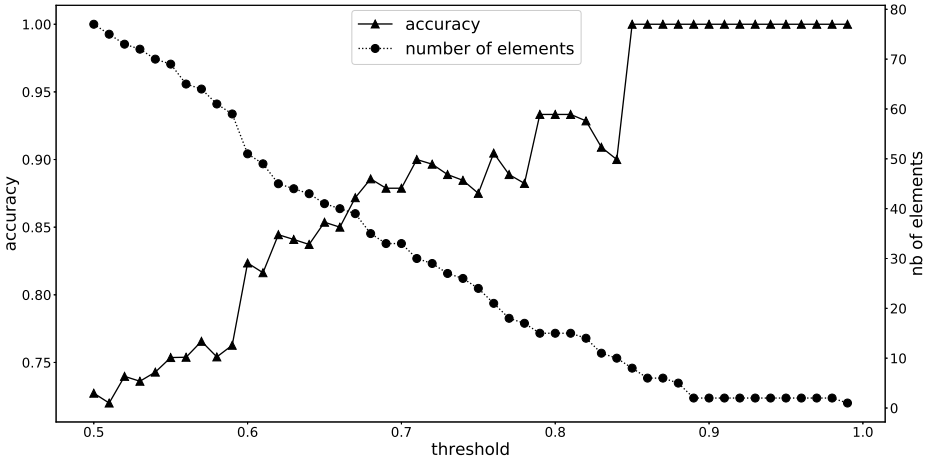

Fig. 6. Accuracy and number of elements according to the threshold t for WormsTwoClass data set: $\mu = 0.3$ and strategy 3

of the labelled instances over-perform strategies that does not take in account such information on uncertainty.

As perspectives we propose to modify the classification part of F-BOSS and F-DTW in order to attribute a weight on the neighbors depending on the distance to the object to predict. This strategy, inspired by some soft k-NN algorithms for non time series data sets, should improve the performances by giving less importance to labeled instances far and uncertain.

Another perspective consists in adapting the soft algorithms to possibilistic labels. Indeed, the possibilistic labels are more suitable for real applications as it allows an expert to assign a degree of uncertainty on an object to a class independently from the other classes. For instance, in a dairy cows application where the goal is to detect anomalies like diseases or estrus [23], the possibilistic labels are simple to retrieve and well appropriated because a cow can have two or more anomalies at the same time (e.g. a diseases and an estrus).

References

1. Bagnall, A., Lines, J., Bostrom, A., Large, J., Keogh, E.: 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)
2. Bernal, J.L., Cummins, S., Gasparrini, A.: Interrupted time series regression for the evaluation of public health interventions: a tutorial. International journal of epidemiology **46**(1), 348–355 (2017)
3. 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)
4. Dau, H.A., Keogh, E., Kamgar, K., Yeh, C.C.M., Zhu, Y., Gharghabi, S., Ratanamahatana, C.A., Yanping, Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G., Hexagon-ML: The ucr time series classification archive (October 2018), `https://www.cs.ucr.edu/~eamonn/time_series_data_2018/`
5. Derrac, J., García, S., Herrera, F.: Fuzzy nearest neighbor algorithms: Taxonomy, experimental analysis and prospects. Information Sciences **260**, 98–119 (2014)
6. Destercke, S.: A k-nearest neighbours method based on imprecise probabilities. Soft Computing **16**(5), 833–844 (2012)
7. El Gayar, N., Schwenker, F., Palm, G.: A study of the robustness of knn classifiers trained using soft labels. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition. pp. 67–80. Springer (2006)
8. Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Deep learning for time series classification: a review. Data Mining and Knowledge Discovery **33**(4), 917–963 (2019)
9. Feyrer, J.: Trade and income—exploiting time series in geography. American Economic Journal: Applied Economics **11**(4), 1–35 (2019)
10. Hüllermeier, E.: Possibilistic instance-based learning. Artificial Intelligence **148**(1-2), 335–383 (2003)
11. Keller, J.M., Gray, M.R., Givens, J.A.: A fuzzy k-nearest neighbor algorithm. IEEE transactions on systems, man, and cybernetics pp. 580–585 (1985)
12. Keogh, E., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. Knowledge and information systems **7**(3), 358–386 (2005)

13. Machanje, D., Orero, J., Marsala, C.: A 2d-approach towards the detection of distress using fuzzy k-nearest neighbor. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. pp. 762–773. Springer (2018)
14. Östermark, R.: A fuzzy vector valued knn-algorithm for automatic outlier detection. Applied Soft Computing **9**(4), 1263–1272 (2009)
15. Quost, B., Denœux, T., Li, S.: Parametric classification with soft labels using the evidential em algorithm: linear discriminant analysis versus logistic regression. Advances in Data Analysis and Classification **11**(4), 659–690 (2017)
16. Ratanamahatana, C.A., Keogh, E.: Three myths about dynamic time warping data mining. In: Proceedings of the 2005 SIAM International Conference on Data Mining. pp. 506–510. SIAM (2005)
17. 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)
18. 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)
19. Schäfer, P., Högqvist, M.: Sfa: a symbolic fourier approximation and index for similarity search in high dimensional datasets. In: Proceedings of the 15th International Conference on Extending Database Technology. pp. 516–527. ACM (2012)
20. Susto, G.A., Cenedese, A., Terzi, M.: Time-series classification methods: Review and applications to power systems data. In: Big data application in power systems, pp. 179–220. Elsevier (2018)
21. Thiel, C.: Classification on soft labels is robust against label noise. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. pp. 65–73. Springer (2008)
22. Tiwari, A.K., Srivastava, R.: An efficient approach for prediction of nuclear receptor and their subfamilies based on fuzzy k-nearest neighbor with maximum relevance minimum redundancy. Proceedings of the National Academy of Sciences, India Section A: Physical Sciences **88**(1), 129–136 (2018)
23. Wagner, N., Antoine, V., Mialon, M.M., Lardy, R., Silberberg, M., Koko, J., Veissier, I.: Machine learning to detect behavioural anomalies in dairy cows under subacute ruminal acidosis. Computers and Electronics in Agriculture **170**, 105233 (2020). https://doi.org/https://doi.org/10.1016/j.compag.2020.105233, `http://www.sciencedirect.com/science/article/pii/S0168169919314905`
24. Zhang, Y., Chen, J., Fang, Q., Ye, Z.: Fault analysis and prediction of transmission line based on fuzzy k-nearest neighbor algorithm. In: 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). pp. 894–899. IEEE (2016)