Evaluation Protocol of Early Classifiers Over Multiple Data Sets

Asma Dachraoui[1][2], Alexis Bondu[1], and Antoine Cornuejols[2]

[1] EDF R&D, 1 avenue du Général de Gaulle, 92140 Clamart, France
[2] AgroParisTech - INRA, UMR-518-MIA, 16 rue Claude Bernard,
F-75005 Paris, France.
{asma.dachraoui,alexis.bondu}@edf.fr
antoine.cornuejols@agroparistech.fr

Abstract. Early classification approaches deal with the problem of reliably labeling incomplete time series as soon as possible given a level of confidence. While developing new approaches for this problem has been getting increasing attention recently, their evaluation are still not thoroughly considered. In this article, we propose a new evaluation protocol for early classifiers. This protocol is generic and does not depend on the criteria used to evaluate the classifiers. We experimented this protocol over multiple data sets.

Keywords: Early classification, Time series, Evaluation protocol

1 Introduction

In recent years, the interest in adapting supervised and unsupervised machine learning (ML) approaches to time series analysis has considerably increased. A part of these studies focuses on applying these approaches on incomplete time series, when they are progressively recorded. This article focuses on the particular context of early classification of time series. Early classification techniques are useful for many time-critical applications. For example, in medicine, earliest diagnosis based on first signal outputs may help to remedy some diseases [1]. In air, road or marine traffic it is important to anticipate the risks of collision or crash before receiving all signals [2]. These approaches allow one to make early and reliable predictions based on incomplete time series.

The conventional time series classification problem consists in predicting the labels of **complete** time series. In this case, the classifier is learned on a training set composed by complete time series. Then the classifier can be applied on new complete time series. By contrast, an early classifier can be applied on **incomplete** time series after being learned on complete time series. In this case, the objective is to predict the labels **as soon as possible** given a level of **confidence**. Early classification can be considered as a multi-objective problem. On the one hand, the objective of **quality** consists in reliably predicting the labels in order to make an **appropriate** action. On the other hand, the objective of

earliness aims to make this action as soon as possible, before a deadline. At last, the early classification problem involves two conflicting objectives.

The evaluation of conventional classifiers has been well studied in the literature. The comparison of two classifiers over multiple data sets requires the use of a statistical test, in order to **objectively** and **reliably** decide if one classifier is better than the other. In [3] the authors recommend to use the Wilcoxon signed-rank test. The comparison of several early classifiers is a special case since it considers both quality and earliness objectives. Therefore, an approach could be better than another on specific objective and worse on the other. In the literature, the methods that optimize several objectives are mainly exploited for learning purposes. For instance, the scalarized multi-objective learning approach which aggregates two objectives into a single scalar objective function allows one to optimize the two objectives by setting a regularization parameter. Multi-Objective Evolutionary Algorithms are used to design classifiers in diverse problems with more than one objective [4]. The majority of these approaches involve one or more regularization parameters which are adjusted by a learning algorithm. These approaches are not suitable for the evaluation.

In this article, we propose a new evaluation protocol for early classifiers. This protocol is based on the Wilcoxon signed-rank test and the Pareto optimum. The proposed protocol is parameter-free and generic since it does not involve an evaluation criterion that "*mixes*" the two conflicting objectives. The remainder of this article is structured as follows: in Sect.2, we present a brief review of early classification of time series. In Sect.3 we propose a new evaluation protocol for early classifiers. The fourth section presents our choices for evaluating the quality and the earliness of the classifiers. The experimental results are discussed in Sect.5. Section 6 concludes this article and highlights the future works.

2 Early Classification

We define a time series $x = \{(t_1, x_1), (t_2, x_2), ..., (t_L, x_L)\}$ of length L as a sequence of real values $\{x_{j \in [1,L]}\}$ associated with the timestamps $\{t_{j \in [1,L]}\}$. The input data set is a collection of N pairs $\{(x_i, y_i) | i \in [1, N]\}$ where x_i is the i^{th} time series, $y_i \in \mathcal{Y}$ is its class value. $\mathcal{Y}$ is a finite set of class labels. By contrast with the conventional classification, the early classification consists in predicting the labels before the time series are completed. In the conventional case, the single objective is to maximize the quality of the prediction without considering the time constraint. In the case of early classification, there are two conflicting objectives to optimize: i) predict the labels of incomplete time series **as soon as possible**; ii) maximize the **quality** of the classifier. In practice, the predictions are triggered once a given level of **confidence** is reached.

Let $\tilde{x}_i$ be an incomplete time series, $H(.)$ is a prediction function which predicts the class label of $\tilde{x}_i$ (*see Eq.1*) and $C(.)$ is a function which measures the confidence of this prediction (*see Eq.2*).

$$H(\tilde{x}_i) \longrightarrow \hat{y} \quad (1) \qquad C_H(\tilde{x}_i) \longrightarrow \sigma \quad (2) \qquad \tilde{t}_i : C_H(\tilde{x}_i) \geq \tau \quad (3)$$

The earliness of a classifier is evaluated based on the instants $\tilde{t}_i$ when the predictions are triggered. More formally, $\tilde{t}_i$ is the earliest timestamp such that the class label of $\tilde{x}_i$ is predicted with a confidence level exceeding a threshold τ (*see Eq.3*).

In the literature, the existing early classification approaches may be distinguished by the manner of setting the confidence. The confidence measures can be decomposed in two categories according to the type of the classifier:

1. The *generative classifiers* provide conditional probabilities of the class values on which are based the confidence measures. The simplest approach consists in triggering the predictions once the probability of $\hat{y}$ exceeds a fixed **threshold** [5][6][7]. This approach is improved in [8] by introducing the concept of the *reliability*. The key idea is to model the missing values of the time series as a random variable, conditionally to the observed data points. For a given $\tilde{x}_i$, the distribution of the possible ways to complete the time series is estimated. The confidence is then evaluated by the part of this distribution leading to the predicted class value.
2. The *discriminative classifiers* provide a class label based on a decision boundary. In this case, the confidence measures are defined using the distance from the decision boundaries [9].

Methods such as [10] implicitly introduce the notion of the confidence by respecting two properties: the consistency and the stability of the predictions. The *Minimum Prediction Length* (MPL) is proposed in order to determine the earliest instant from which the prediction will be the same as if the full-length time series is used. An extended *1-Nearest Neighbor* (1NN) approach is proposed.

3 Evaluation over multiple data sets

The rest of this article focuses on the evaluation of early classifiers. In this section, we first discuss existing research work on comparing conventional classifiers. Thereafter, we propose a new protocol to compare early classifiers.

3.1 Comparison of two classifiers

The evaluation of conventional classifiers are thoroughly considered in the literature. Numerous evaluation criteria such as the accuracy, the *Balanced error Rate* (BER) or the *Area Under the ROC Curve* (AUC) allow one to evaluate the classifiers on a specific data set. According to J. Demšar [3], a reliable evaluation of several classifiers should be done over multiple data sets, in order to reflect the general quality and not the quality on a specific data set. If the output value *(denoted by z)* is smaller than -1.96, the quality of both classifiers are significantly different.

3.2 A new evaluation protocol for early classification

Our objective is to **reliably** compare two early classifiers. In this section, we propose a generic evaluation protocol which does not depend on the criteria used to evaluate the quality and the earliness of the classifiers.

Step 1 - Prediction of the labels over multiple data sets :
Let $\mathcal{D}$ be an ensemble of K data sets denoted by $\mathcal{D} = \{D_1, D_2, \ldots, D_K\}$. Each data set is divided into two disjoint training set and test set. Let $H_{(A)}$ and $H_{(B)}$ be the prediction functions associated with two early classifiers denoted by A and B. Both classifiers are trained over the **same** training sets of complete time series. For each test set, two pairs of scores $[Q(.), T(.)]$ are computed for the classifiers A and B. The criterion $Q(.)$ evaluates the quality of the predictions and $T(.)$ corresponds to the earliness. $Q(.)$ and $T(.)$ are two global criteria which are computed by processing all the time series of a given test set. Thus, these time series are processed one by one at each instant. A confidence measure denoted by $C_H(.)$ is compared with a fixed threshold τ in order to trigger the predictions. If $C_H(.)$ does not exceed τ before the end of the time series, the prediction is triggered at the last timestamp. As shown in Fig.1, the instants of the predictions vary for each time series. During this process, the following pieces of information are retained for each time series $\tilde{x}_i$: y_i the true label, $\hat{y}_i$ the predicted class value, $\tilde{t}_i$ the instant of the prediction and $\hat{P}_i(y_i, \tilde{x}_i)$ the predicted probabilities of the class values. These pieces of information, denoted by $\mathcal{I}$, are used to compute the values of $Q(.)$ and $T(.)$ over the entire test set.

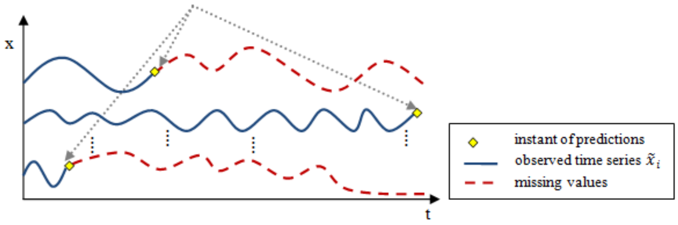

Fig. 1. Illustration of the labels predictions triggering.

Step 2 - Independent comparison under each objective :
The criteria $Q(.)$ and $T(.)$ are computed for each data set and for each classifier. Then, the Wilcoxon signed-rank test is independently performed for the both objectives: the quality and the earliness. Based on the statistic z values, we can make conclusions about which classifier is the best either under the quality objective or under the earliness objective.

Step 3 - Global comparison by using Pareto optimum :
In order to compare two early classifiers, we drawn inspiration from the Pareto optimum: a classifier is considered as better than an other if it improves at least one of the two objectives without degrading the other. Based on the values of the

statistic z independently computed for the quality and the earliness objectives, there are three possible cases: *i)* A is better than B, *ii)* B is better than A and *iii)* A and B are indiscernible since there is no significant differences under the two objectives.

4 Evaluation criteria for early classifiers

This section presents a possible implementation for the quality, the earliness and the confidence measures.

4.1 A measure for prediction quality

The evaluation of the quality of the classifiers can be measured by diverse criteria. We choose to exploit the Multi-class Area Under the ROC Curve (AUC) criterion [11]. For a given class label y_j, the area under the ROC curve AUC_{y_j} relatively to y_j is computed based on the triggered predictions (*see Fig.1*):

$$Q(.) = \mathbb{E}_{j=1}^{J}[AUC_{y_j}] = \sum_{j=1}^{J} P_k(y_j) AUC_{y_j} \tag{4}$$

4.2 A measure for earliness

The earliness measure that we propose allows us to quantify the earliness of an early classifier over all the time series in the test set. This measure is computed by exploiting the set of information $\mathcal{I}$ (*see Sect.3.2*). We define:

1. **the time dimension:** $t = \frac{\tilde{t}_i}{L}$ as the proportion of the total length L (*with L is the length of a complete time series*).

2. **the earliness:** $Pr(t) = \frac{n_t}{N}$ as the proportion of the triggered predictions at instant t (*with n_t is the number of the triggered predictions at the instant t, and N is the size of the test set*).

We denote $Pr(t)$ (*with* $t \in [0,1]$) as the Earliness Curve of the classifier over the time dimension. The global earliness of the classifier (*denoted by* $T(.)$) is measured by the Area Under the Earliness Curve (AUEC), with $T(.) \in [0,1]$ since $t \in [0,1]$ and $Pr(t) \in [0,1]$. When the area $T(.)$ is equal to 1.0, that would mean that all the predictions are triggered at the first instant. In this case the classifier performs a perfect earliness. A $T(.)$ of 0.0 reports that the classifier is not early. In this case, the classifier is too conservative and prefers to wait until the last instant to make the predictions. At last, a $T(.)$ equal to 0.5 means that the average number of triggered predictions at each instant is equals to $\frac{N}{L}$.

4.3 A measure for confidence

In this section, we propose a possible way to fix the confidence level. The prediction of the label of the time series $\tilde{x}_i$ is triggered once the probability of the predicted class $(denoted\ by\ max_1(\hat{P}(y|\tilde{x}_i))$ exceeds k times the probability of the second most probable class $(denoted\ by\ max_2(\hat{P}(y|\tilde{x}_i))$:

$$\tilde{t}_i : max_1(\hat{P}(y|\tilde{x}_i)) \geq k * max_2(\hat{P}(y|\tilde{x}_i)) \tag{5}$$

5 Experiments

The objective of our experiments is to compare two early classifiers. In this section, we exploit the generic evaluation protocol proposed in Sect.3.2.

Implementation of the early classifiers:
In this article, early classification is implemented based on a collection of classifiers trained in parallel [12]. Each classifier corresponds to one instant of the time series. These classifiers do not exploit the same explicative variables. The progressive arrival of the time series is simulated by hiding the forthcoming data points: the input of the current classifier is only composed by the previous data points up to the current instant. Fig.2 illustrates this implementation.

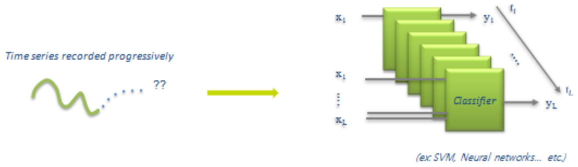
Time series recorded progressively
??
x_1
x_1
x_L
y_1
t_1
Classifier
y_L
t_L
(ex: SVM, Neural networks... etc.)

Fig. 2. Implementation of early classification over a collection of classifiers.

In order to use the proposed evaluation protocol, we implement two different classifiers based on the above described implementation: *i)* a *Selective Naive Bayes* (SNB) using a the regularized approach MODL [13] which shows a capacity for a good discrimination between classes and *ii)* a *Naive Bayes* (NB) with 10 *EqualFrequency* as a baseline for comparison.

Data sets description:
Our experiments were performed over 23 data sets selected from the UCR Time Series Classification and Clustering repository [14]. We randomly and disjointedly re-sampled each data set into a set of 70% of examples for the training set and the remainder for the test set.

Results:
In this paragraph, we evaluate our generic protocol by experimenting the classifiers under several data sets. We first run the SNB and the NB classifiers over all the data sets in order to compute the quality and the earliness measures. Then,

we applied the Wilcoxon test for the two classifiers independently under the quality and the earliness objectives. The results are listed in Tab.1. The quality of the predictions $Q(.)$ and the earliness $T(.)$ results of each classifier over each data set are depicted with a varying value of k (*see Eq.5*). In our experiments, we varied k from 2 to 12 with a step equal to 1. But for simplicity, Tab.1 shows only the results for k ranging from 2 to 10 with a step equal to 4 (*the omitted results do not affect the drawn conclusions*). The last line in the table represents the values of the statistic z for overall predictions quality and earliness results. Significant values are reported in bold font.

Table 1. Empirical results of the SNB and NB early classifiers.

	Prediction quality						**Earliness**					
k	**2**		**6**		**10**		**2**		**6**		**10**	
Data Set	$\mathbf{Q}_{SNB}$	$\mathbf{Q}_{NB}$	$\mathbf{Q}_{SNB}$	$\mathbf{Q}_{NB}$	$\mathbf{Q}_{SNB}$	$\mathbf{Q}_{NB}$	$\mathbf{T}_{SNB}$	$\mathbf{T}_{NB}$	$\mathbf{T}_{SNB}$	$\mathbf{T}_{NB}$	$\mathbf{T}_{SNB}$	$\mathbf{T}_{NB}$
50words	0.505	0.527	0.520	0.530	0.519	0.527	0.766	0.687	0.679	0.564	0.642	0.513
CBF	0.807	0.840	0.866	0.875	0.877	0.890	0.803	0.814	0.717	0.739	0.697	0.722
ChlorineC	0.5	0.549	0.764	0.648	0.799	0.655	0.993	0.988	0.6104	0.667	0.455	0.595
CinC	0.777	0.669	0.8007	0.677	0.854	0.699	0.758	0.988	0.706	0.977	0.622	0.971
CricketX	0.468	0.488	0.475	0.481	0.478	0.482	0.856	0.992	0.807	0.990	0.799	0.990
CricketY	0.523	0.478	0.515	0.470	0.496	0.470	0.985	0.927	0.956	0.920	0.933	0.920
CricketZ	0.483	0.514	0.491	0.511	0.489	0.510	0.95	0.376	0.922	0.339	0.903	0.339
ECG5Days	0.955	0.793	0.992	0.824	0.994	0.839	0.318	0.614	0.209	0.580	0.200	0.568
FaceAll	0.483	0.476	0.489	0.467	0.482	0.475	0.940	0.878	0.809	0.738	0.748	0.687
FaceUCR	0.503	0.536	0.496	0.524	0.497	0.533	0.969	0.989	0.959	0.969	0.938	0.941
Mallat	0.910	0.928	0.913	0.930	0.916	0.931	0.886	0.718	0.845	0.709	0.833	0.709
MedicalImg	0.603	0.529	0.587	0.520	0.589	0.519	0.733	0.560	0.324	0.475	0.180	0.424
MoteStain	0.994	0.962	0.994	0.962	0.994	0.962	0.006	0.085	0.006	0.084	0.006	0.084
StarLight	0.742	0.920	0.738	0.919	0.734	0.919	0.988	0.715	0.996	0.683	0.996	0.716
SwedishLeaf	0.539	0.484	0.527	0.482	0.530	0.480	0.942	0.742	0.919	0.721	0.903	0.720
Symbols	0.958	0.934	0.958	0.935	0.958	0.940	0.610	0.815	0.609	0.811	0.609	0.818
TwoLECG	0.983	0.921	0.986	0.936	0.987	0.945	0.3539	0.464	0.341	0.424	0.336	0.393
uWaveX	0.719	0.702	0.724	0.701	0.734	0.692	0.966	0.993	0.965	0.993	0.930	0.991
uWaveY	0.748	0.787	0.749	0.796	0.749	0.796	0.992	0.386	0.988	0.390	0.987	0.390
uWaveZ	0.775	0.727	0.780	0.726	0.781	0.723	0.751	0.799	0.728	0.797	0.726	0.795
wafer	0.958	0.925	0.981	0.924	0.98	0.940	0.923	0.909	0.889	0.840	0.883	0.824
WordsSyn	0.54	0.518	0.541	0.509	0.545	0.509	0.928	0.786	0.654	0.647	0.630	0.647
yoga	0.764	0.658	0.889	0.640	0.907	0.693	0.700	0.795	0.407	0.761	0.308	0.748
StatisticZ	-1.3686729		**-2.28112149**		**-3.10232523**		-0.36497944		-0.36497944		-0.69954392	

Based on the values of the statistic z and the Pareto optimum, we can conclude that SNB is better than NB under the two objectives. In fact, by varying k we only observe two cases: *1)* SNB and NB are indiscernible and *2)* SNB is better than NB. For example for $k = 2$, the two classifiers are indiscernible. For $k = 6, 10$, SNB is always better than NB under the quality objective and SNB and NB are always indiscernible under the earliness objective. At the end, the following results show that the proposed protocol performs as expected: the SNB classifier outperforms the NB baseline classifier.

6 Conclusion

In this article, we proposed a new evaluation protocol to compare two early classifiers under two conflicting objectives: the quality and the earliness. The protocol is generic and does not depend on the implementation choices made for computing the two objectives. Overall, the results suggest that this protocol could represent a useful evaluation technique to objectively and reliably compare early classifiers. Further, a limitation that we identified indicates that the Wilcoxon test does not ensure the concomitance of the results for all the data sets under the two objectives. It is possible that our protocol suggests that a classifier is better than another under both objectives without this may be true for the same data set. This is due to the fact that the Wilcoxon test is based on the differences over the two evaluation criteria. Thus, we only guarantee that the largest differences over the two criteria are observed for the same classifier. As future work, we intend to build a benchmark of early classification approaches in the literature based on our proposed protocol.

References

1. S. G. Nayak, O. Davide, and C. Puttamadappa, "Classification of bio optical signals using k- means clustering for detection of skin pathology," *International Journal of Computer Applications (IJCA)*, vol. 1, no. 2, pp. 112–116, (2010).
2. J.-M. Chiou, "Dynamical functional prediction and classification, with application to traffic flow prediction.," *The Annals of Applied Statistics*, vol. 6, no. 4, pp. 1588–1614, (2012).
3. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. (2006).
4. Y. Jin and B. Sendhoff, "Pareto-based multiobjective machine learning: An overview and case studies.," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 38, no. 3, pp. 397–415, (2008).
5. N. Hatami and C. Chira, "Classifiers With a Reject Option for Early Time-Series Classification," *CoRR*, (2013).
6. K. Trapeznikov and V. Saligrama, "Supervised Sequential Classification Under Budget Constraints," in *AISTATS*, pp. 581–589, (2013).
7. H. S. Anderson, N. Parrish, K. Tsukida, and M. R. Gupta, "Reliable early classification of time series," in *ICASSP*, pp. 2073–2076, (2012).
8. H. S. Anderson, N. Parrish, K. Tsukida, and M. R. Gupta, "Early Time-Series Classification with Reliability Guarantee," tech. rep., SANDIA Laboratories, (2012).
9. S. J. Delany, P. Cunningham, D. Doyle, and A. Zamolotskikh, "Generating estimates of classification confidence for a case-based spam filter.," in *ICCBR*, vol. 3620 of *Lecture Notes in Computer Science*, pp. 177–190, Springer, 2005.
10. Z. Xing, J. Pei, and P. S. Yu, "Early prediction on time series: A nearest neighbor approach.," in *IJCAI* (C. Boutilier, ed.), pp. 1297–1302, (2009).
11. A. Bondu, *Active Learning using Local Models.* PhD thesis, University of Angers, (2008).
12. A. Dachraoui, A. Bondu, and A. Cornuejols, "Early classification of individual electricity consumptions," *RealStream2013 (ECML)*, pp. 18–21, (2013).
13. M. Boullé, "Data grid models for preparation and modeling in supervised learning," in *Hands-On Pattern Recognition: Challenges in Machine Learning*, vol. 1, pp. 99–130, Microtome, (2011).
14. E. Keogh, X. Xi, L. Wei, and C. Ratanamahatana, "The UCR Time Series Classification/Clustering Homepage," (2006).