An Iterative Feature Filter for Sensor Timeseries in Pervasive Computing Applications

Davide Bacciu

Dipartimento di Informatica, Università di Pisa, Italy
bacciu@di.unipi.it

Abstract. The paper discusses an efficient feature selection approach for multivariate timeseries of heterogeneous sensor data within a pervasive computing scenario. An iterative filtering procedure is devised to reduce information redundancy measured in terms of timeseries cross-correlation. The algorithm is capable of identifying non-redundant sensor sources in an unsupervised fashion even in presence of a large proportion of noisy features. A comparative experimental analysis on real-world data from pervasive computing applications is provided, showing that the algorithm addresses major limitations of unsupervised filters in literature when dealing with sensor timeseries.

1 Introduction

Pervasive computing puts forward a vision of an environment enriched by a distributed network of devices with heterogeneous sensing and computational capabilities, that are used to realize customized services supporting everyday activities. Pervasive computing systems deploy sensors that continuously collect data concerning the user and/or the environmental status. This data comes under the form of streams, i.e. timeseries, of sensor information with a considerably heterogeneous nature (e.g. temperature, presence, motion, etc.). This results in consistent amounts of information that need to be transferred and processed, typically in real time, to implement the system services, that are often realized by computational learning models (e.g. for predicting user activities based on sensed data) [1]. In this context, feature selection techniques for multivariate timeseries are fundamental, on one hand, to reduce the computational and communication overhead of transferring and processing such large amounts of sensor information. On the other hand, they serve to suppress redundant/irrelevant information which might negatively affect the predictive performance of the learning model.

We consider a pervasive learning system realized as part of the RUBICON project [2], that consists of a network of learning modules distributed on sensor motes characterized by limited computational and communication capabilities. Each of such device hosts a learning component implementing an Echo State Network (ESN) [3] model which is trained to perform real-time predictive tasks based on the data gathered by the sensors onboard the mote or received from another node through its radio interface. The optimization of the number of sensor streams feeding the learning modules is a key issue in such a resource constrained environment, requiring effective feature selection techniques

V. Mladenov et al. (Eds.): EANN 2014, CCIS 459, pp. 39–48, 2014.

for multivariate timeseries. Further, the RUBICON learning system allows to incrementally deploy new predictive tasks during system's operation, posing additional requirements on the feature selection model. The first is computational efficiency, as the selection process has to be performed in during system operation whenever a request for a new predictive task is posted. The second is the automatization of the feature selection process, as this has to be performed automatically by the learning system without any form of human/expert intervention (e.g to determine the number of selected features from a ranking).

Feature selection entails the identification of a subset of the original input sequences from a given dataset targeted at removing irrelevant and/or redundant information sources. In literature, the majority of the feature selection algorithms for multivariate timeseries take a wrapper approach, where the feature subset is selected to optimize the predictive and generalization abilities of a specific computational learning model [4]. Wrapper approaches are characterized by considerable computational requirements due to the burden of the multiple retraining of the underlying learning model and their results cannot be generalized to a different learning model. Corona (Correlation as Features) [5], for instance, is a wrapper method that transforms each multivariate timeseries into the corresponding correlation matrix, whose coefficients are fed to a support vector machine that is then used to apply the Recursive Feature Elimination method by [4]. Filter approaches, instead, use an external optimization criterion with respect to the learning model that will be using the selected data. Most of the filter techniques for timeseries data are tailored to classification tasks [6], as they select the streams that best separate multivariate samples from different classes. The Relief method, originally proposed for vectorial data, uses entropy as a measure of the ability of a feature to discriminate classes and has been extended to timeseries data [7]. The CleVer method [8] is one of the few unsupervised filter approaches for multivariate timeseries: it exploits the properties of the principal components common to all the timeseries to provide a ranking of the more informative features. Based on the assumption that there exists a common subspace across all multivariate data items, it first performs PCA on each univariate timeseries and then obtains the common principal components by bisecting the angles between their principal components. The CleVer method has found wide application, mainly due to its low computational requirements. However the number of selected features is not determined by the algorithm, rather it is selected by the user (as in k-means).

Previous works have noted how such sophisticated state-of-the-art feature selection techniques, which show excellent performances on multivariate timeseries benchmarks, do not provide significant results in context of open-ended discovery in real-world scenarios comprising a sensor-rich environment [9]. Motivated by this, we propose a simple, yet effective, feature selection technique based on a cross-correlation analysis of multivariate sensor timeseries, that is specifically tailored to the identification and removal of redundant sensor streams in an autonomous fashion. The proposed approach is based on an iterative filter heuristics that incrementally removes/selects timeseries based on redundancy information.