Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 114 (2017) 258–265

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS
October 30 – November 1, 2017, Chicago, Illinois, USA

Interactive Pattern Discovery in High-Dimensional, Multimodal Data Using Manifolds

Jinhong K. Guo and Martin O. Hofmann

Lockheed Martin Advanced Technology Laboratories
3 Executive Campus, Suite 600, Cherry Hill, NJ 08002

Abstract

Data mining to discover patterns and aid decisions is the key to utilizing massive data for process automation and optimization. An especially challenging data mining problem is kriging, i.e., prediction of multiple, related variables from latent patterns in the data. We present a manifold based machine learning approach to discover patterns in massive, correlated, high-dimensional data. Dimensionality reduction using a manifold is a type of non-linear principal component analysis (PCA). The manifold captures the underlying data structure of the inputs and corresponding outputs by way of projecting the data onto a set of basis functions defined by the manifold. These bases ensure that any future adjustments affect the model with respect to the natural geometry of the data. We chose the manifold learning technique for its robustness against unbalanced data. Our contribution, described in this paper, enables interactive learning and incremental learning, i.e., incremental adjustment of the manifold (and its predictions) based on new observations and also user corrections to the predicted values, rerun the analysis on the full data set. Our experiments demonstrate that prediction performance remains equivalent to Multi-kernel Gaussian Processes on standard data sets despite these practically useful enhancements.

Peer-review under responsibility of the scientific committee of the Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems.

Keywords: Machine Learning; Dimensionality Reduction; Data Analysis; Interactive Learning; Incremental Learning; Nyström extension

1. Introduction

Learning and estimating multiple correlating values has been maturing over the years, e.g., Gaussian Processes (GPs) have become one of the mainstream tools for this machine learning application. Though GPs were usually formulated for a single output, researchers have developed multiple output GPs [15]. This enables learning and

Peer-review under responsibility of the scientific committee of the Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems.
10.1016/j.procs.2017.09.037

discovering patterns from high volume of multi-dimensional, incomplete data in many applications and provide decision support.

For many decision support applications it is desirable to allow human experts to interact with, tune, and re-align the system with changing conditions, i.e., to improve or maintain system performance over time and external change. Another desirable feature for a machine learning system is incremental adaptation to new data as they are being discovered. For many approaches, including most neural networks and multi-kernel Gaussian processes, both user corrections and arrival of new data require to completely retrain the underlying model, while our manifold method admits incremental changes.

In this paper, we describe a manifold learning based approach to learn and estimate multiple correlated values. Experiments show that our manifold learning based methods achieve similar results as multi-kernel Gaussian processes. A unique feature of the manifold learning method is that it facilitates incremental and interactive learning which allows gradual model adaptation from new data and from user feedback. We build an initial model using training data. The initial model is optimal at the time of its creation. However, changes, environmental or otherwise, require modification to the model. Our interactive learning technique provides a means for the user and/or feedback to adjust the model over time and for the system to use each adjustment to the maximum effect.

2. Related Work

Similar to active learning, interactive learning seeks to improve learning through labels or adjustments. Interactive learning is more user-driven as opposed to active learning's full automation. The key factor to consider in interactive learning is the amount of machine learning expertise that is required of the user. Some interactive learning systems let users interact directly with the internals of the model. This type of approach enables fine control over the resulting model. However, it requires high level of understanding of the machine learning technique embedded. For example, user can control each split in decision tree for decision tree construction [11]. Our approach let the user treats the modeling process as a black box and interact only with the model's output. Fails and Olsen [12] applies the same philosophy in designing object recognition in images. In their system, users use rough sketches of objects and correct the resulting recognition boundaries based on the metaphor of coloring with crayons. desJardins, et. al., [13] let users move data instances in a spring-embedding layout to correct the model output in visual clustering. The system, based on constrained clustering, infers the constraints based on the users' adjustments.

Our approach is an extension to an earlier research [2] performed at Lockheed Martin Advanced Technology Laboratories. The system interactively refining an initial scoring function. The user first manually specifies a simple linear model as the starting point for learning. Users then examine the output of the model in the 2D scatterplot, and iteratively correct the model by repositioning data instances to the correct level. The system incorporates each correction into the learning process, using the manifold geometry underlying the data to determine the extent of each correction's effect on the learned model. The system targets applications that a given pool of data is consistently being monitored as the scores being adjusted only on existing data instances. Our extension enables us to predict scores for new data instances and allows adjustment to be made using scores on new data instances. This not only facilitate interactive learning that enables user interact with the system to enhance the internal model without any knowledge of it; but also enables incremental learning that incorporates new instances to enrich the model as new data emerges. The incremental characteristic allows us to use a fraction of data for training and incrementally adding the rest to enhance the model, scaling up to large data sets.

Traditionally, statistical regression has been used for value estimation. However, the technique is vulnerable when applied to high dimensional space, when attributes interact, and/or data are unbalanced. Traditional Bayesian approaches require a priori distribution which may be difficult to quantify for some applications. Neural Networks, being used on regression problems [7] needs to determine a large number of parameters from large amounts of training data and have poor scalability. Over the past years, Gaussian Processes have been used as a Bayesian prior over functions and can be viewed as Bayesian linear regression with infinite number of basis functions [8][9] [10]. These Gaussian Processes inference is mostly formulated for a single value. Applications, such as geophysical exploration, require infer multiple, interacting values jointly, preferably using the dependencies to improve the results. Melkumyan and Ramos [15] generalized the multi-task Gaussian process to allow the use of multiple covariance functions, possibly having a different covariance function per task.