TEST (2020) 29:38–41
https://doi.org/10.1007/s11749-019-00696-w

DISCUSSION

Comments on: On Active Learning Methods for Manifold Data

Mostafa Reisi Gahrooei[1] · Hao Yan[2] · Kamran Paynabar[3]

Published online: 2 January 2020

1 Introduction

Learning low-dimensional structures from high-dimensional data, as discussed by Li, Del Castillo, and Runger, is an important area of research with numerous application opportunities. This paper presents an interesting idea of combining Gaussian process classification (GPC) with manifold and active learning. Gaussian processes are strong tools for active learning as they can quantify the model uncertainty of unlabeled samples, while considering the spatial information of input data. However, regular GP models and their kernels are not able to model the manifold on which the data lie. In order to address this issue, the authors take the well-known graph representation of manifolds and define a new kernel by using the graph's Laplacian matrix that captures the manifold structure. Defining such a manifold-based kernel is an important contribution because with this kernel directly inherits all the strengths of the GP models and can be effectively employed for tackling the active learning problem on manifolds.

The proposed approach has broad applications in manufacturing and service systems, biology and genetics, medicine, marketing, internet retails, and so forth. In most applications, the data are adequate, but the labeled data may be quite limited due to the expensive labeling process in different systems. For example, in medical imaging for cancer diagnostics, clinicians need to label each image as benign and malicious, which is a very time-consuming process (Smailagic et al. 2018).

This comment refers to the invited paper available at: https://doi.org/10.1007/s11749-019-00694-y.

✉ Kamran Paynabar
kpaynabar3@gatech.edu; kamran.paynabar@isye.gatech.edu

[1] Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA

[2] School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA

[3] H. Milton School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA

In what follows, we first present two real examples of such applications in manufacturing systems. Next, we discuss potential research directions that could help improve the proposed methodology and broaden its applicability. Finally, concluding remarks are presented.

2 Application examples in manufacturing systems

In modern vehicle engines, the engine control unit (ECU) ensures the functionality of the vehicle and diagnoses failure for a number of components. The ECU implements surrogate models of complex physical dynamical systems that are constructed based on a large number of tests performed at different levels of engine torque and speed. The combinations of the torque and speed that produce acceptable (unacceptable) engine performance may lie on a manifold rather than distributed fully on R^2. Therefore, taking the manifold structure of the data to create a classification model that identifies acceptable and unacceptable pairs of torque and speed is essential. Furthermore, to reduce the cost of performing experiments an active learning approach that identifies the next point to be tested is critical in defining an accurate classification model while minimizing the number of experiments (Gahrooei et al. 2019). The proposed approach in this paper provides a strong solution to achieve this goal.

Another potential application of the proposed GCP is in quality inspection of complex structured parts. For example, in metrology, touch-probe coordinate measuring machines used for measuring the dimensional accuracy (Mesnil et al. 2014), and in non-destructive evaluation, guided wave-field tests and laser ultrasonics are widely utilized for defect detection in composite sheets (Simpson 1992). These inspection systems are only capable of measuring one point at a time, resulting in a time-consuming procedure not scalable to online inspection of complex parts. However, since defects are often clustered, one can use a sequential sampling and active learning strategy such as the proposed GPC to reduce the number of inspected points and, consequently, reduce the inspection time. This approach may help identify the location and shape of the defects without having to inspect every single point on the part.

3 Research directions

In this section, we discuss the some of the assumptions and limitations of the proposed methodology and suggest some research directions for addressing them. Since the proposed method uses Gaussian process (GP) classification, it has the weaknesses of the GP as well. Specifically, the approach is particularly advertised to be useful in high-dimensional settings in which GP models are particularly computationally expensive. It is known that when $d > 10$ or the sample size n is large (more than a few thousands), fitting a GP would be extremely time-consuming. Considering the fact that in an active learning procedure, the GP model should be updated as new samples are being labeled, the whole framework seems to

be intensely time demanding. This suggests several lines of research to be pursued: (1) investigating the viability of existing techniques for improving the time complexity of GP models with graph-based kernels. In recent years, several techniques have been introduced to address the challenge of time complexity in GP models. Examples of these approaches are LaGP, DICE, and sparse Gaussian process (Gramacy 2016; Roustant et al. 2012; Snelson and Ghahramani 2006). Nevertheless, almost all these approaches focus on the large sample size problem rather than high-dimensionality (large d) problem. (2) Developing sequential and recursive estimation of the GP model parameters that does not require retraining of the GP model for entire data at each epoch of the active learning procedure. Although GP models are suitable for active learning as they quantify the model uncertainty at unobserved points, retraining these models at each step of active learning can make them intractable due to their large computation time. Therefore, employing GP models for active learning in high-dimensional settings may not be appealing unless a recursive and fast scheme for updating the GP parameters is devised. (3) Utilizing the manifold structure of the data to reduce the computation effort and time of the parameter estimation. The GP model developed in the paper is defined on a graph constructed by manifold learning, which typically has sparse connections. In the literature, to fit GP models on spatial data, the entire space is divided into several non-overlapping regions and a local GP for each region is fitted, while enforcing the boundary consistency between regions. Similar techniques can be applied to the graph-based GP, where community detection can be used to divide the space into clusters and fit GP for each local cluster, while considering the global continuity.

Another point that is not addressed in the paper is that how the choice of the base kernel influences the performance of the proposed approach in terms of capturing the manifold structure. The proposed kernel indeed fully depends on the base kernel, and hence, its ability in capturing the manifold structure may vary from one base kernel to another. However, kernel selection is a data-specific problem. In the GP literature, there is some research discussing the selection of hyper parameters and also even deciding the functional forms of kernel, which can be further studied.

Another challenge is with regard to the generation of graphs to represent the manifold in a high-dimensional space. The graph representation of a manifold is usually defined by connecting the points that are within ε distance of each other (i.e., an edge will be placed between two points if their Euclidian distance is less than ε in the ambient space). In lower dimensions (e.g., $d = 2$ or 3), this approach is reasonable with small to medium number of data points. However, in higher dimensions, one may require a large number of data points to be able to create such a graph as most points fall far from each other. Therefore, the created graph will be extremely sparse and may not be a proper representative of the underlying manifold. As a result, employing manifold learning techniques may not be straightforward in high-dimensional settings.

Finally, we discuss the use of the deep learning model as an alternative model for manifold learning. Deep learning has shown great success in supervised learning in image recognition and natural language processing (Deng et al. 2009). Recently, deep learning models have been developed in the semi-supervised learning settings

in conjunction with active learning and it has shown promising results (Gal et al. 2017) with entropy-based techniques. The benefit of deep learning approaches is that unlike the graph-based manifold learning that uses Euclidian distance, they do not require a definition of a distance measure. Deep learning methods also map the high-dimensional data into the low-dimensional feature space. Nevertheless, both approaches will require a large number of data points for training a mapping function from the original space to the manifold where the data lie. The flexibility of neural network architecture such as convolutional and recurrent structure makes it more powerful when the number of data increases.

4 Concluding remarks

Li, Del Castillo, and Runger's paper is an exciting contribution to the literature of high-dimensional data analysis. The authors are congratulated for developing such an interesting framework integrating manifold learning, GP classification, and active learning. In addition to the potential manufacturing applications discussed in Sect. 2, there are various other scenarios where sampling is costly and/or time-consuming, hence requiring an active learning and sequential sampling on manifolds. As mentioned in Sect. 3, this research opens the door to a broad range of research problems and challenges that require methodological development as well as theoretical study.

References

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255

Gahrooei MR, Paynabar K, Pacella M, Colosimo BM (2019) An adaptive fused sampling approach of high-accuracy data in the presence of low-accuracy data. IISE Trans 51(11):1–14

Gal Y, Islam R, Ghahramani Z (2017) Deep Bayesian active learning with image data. In: Proceedings of the 34th international conference on machine learning, vol 70, pp 1183–1192

Gramacy RB (2016) laGP: large-scale spatial modeling via local approximate Gaussian processes in R. J Stat Softw 72(1):1–46

Mesnil O, Yan H, Ruzzene M, Paynabar K, Shi J (2014) Frequency domain instantaneous wavenumber estimation for damage quantification in layered plate structures. In: EWSHM-7th European workshop on structural health monitoring, 2014

Roustant O, Ginsbourger D, Deville Y (2012) DiceKriging, DiceOptim: two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization

Simpson JA (1992) Mechanical measurement and manufacturing. Control Dyn Syst Adv Theory Appl 45:17

Smailagic A, Costa P, Noh HY, Walawalkar D, Khandelwal K, Galdran A, Mirshekari M, Fagert J, Xu S, Zhang P, Campilho A. MedAL (2018) Accurate and robust deep active learning for medical image analysis. In: 17th IEEE international conference on machine learning and applications (ICMLA) 2018, pp 481–488

Snelson E, Ghahramani Z (2006) Sparse Gaussian processes using pseudo-inputs. In: Advances in neural information processing systems, pp 1257–1264