Technometrics, ISSN 0040-1706
Volume 64, number 2 (may 2022)

Class Maps for Visualizing Classification Results P. 151-165

Jakob Raymaekers, Peter J. Rousseeuw, Mia Hubert

Abstract

Classification is a major tool of statistics and machine learning. A classification method first processes a training set of objects with given classes (labels), with the goal of afterward assigning new objects to one of these classes. When running the resulting prediction method on the training data or on test data, it can happen that an object is predicted to lie in a class that differs from its given label. This is sometimes called label bias, and raises the question whether the object was mislabeled. The proposed class map reflects the probability that an object belongs to an alternative class, how far it is from the other objects in its given class, and whether some objects lie far from all classes. The goal is to visualize aspects of the classification results to obtain insight in the data. The display is constructed for discriminant analysis, the k-nearest neighbor classifier, support vector machines, logistic regression, and coupling pairwise classifications. It is illustrated on several benchmark datasets, including some about images and texts.

SPlit: An Optimal Method for Data Splitting P. 166-176

V. Roshan Joseph, Akhil Vakayil

Abstract

In this article, we propose an optimal method referred to as SPlit for splitting a dataset into training and testing sets. SPlit is based on the method of support points (SP), which was initially developed for finding the optimal representative points of a continuous distribution. We adapt SP for subsampling from a dataset using a sequential nearest neighbor algorithm. We also extend SP to deal with categorical variables so that SPlit can be applied to both regression and classification problems. The implementation of SPlit on real datasets shows substantial improvement in the worst-case testing performance for several modeling methods compared to the commonly used random splitting procedure.

Bayesian Hierarchical Model for Change Point Detection in Multivariate Sequences P. 177-186

Huaqing Jin, Guosheng Yin, Binhang Yuan & Fei Jiang

Abstract

Motivated by the wind turbine anomaly detection, we propose a Bayesian hierarchical model (BHM) for the mean-change detection in multivariate sequences. By combining the exchange random order distribution induced from the Poisson–Dirichlet process and nonlocal priors, BHM exhibits satisfactory performance for mean-shift detection with multivariate sequences under different error distributions. In particular, BHM yields the smallest detection error compared with other competitive methods considered in the article. We use a local scan procedure to accelerate the computation, while the anomaly locations are determined by maximizing the posterior probability through dynamic programming. We establish consistency of the estimated number and locations of the change points and conduct extensive simulations to evaluate the BHM approach. Among the popular change point detection algorithms, BHM yields the best performance for most of the datasets in terms of the F1 score for the wind turbine anomaly detection.

PICAR: An Efficient Extendable Approach for Fitting Hierarchical Spatial Models

P. 187-198

Ben Seiyon Lee, Murali Haran

Abstract

Hierarchical spatial models are very flexible and popular for a vast array of applications in areas such as ecology, social science, public health, and atmospheric science. It is common to carry out Bayesian inference for these models via Markov chain Monte Carlo (MCMC). Each iteration of the MCMC algorithm is computationally expensive due to costly matrix operations. In addition, the MCMC algorithm needs to be run for more iterations because the strong cross-correlations among the spatial latent variables result in slow mixing Markov chains. To address these computational challenges, we propose a projection-based intrinsic conditional autoregression (PICAR) approach, which is a discretized and dimension-reduced representation of the underlying spatial random field using empirical basis functions on a triangular mesh. Our approach exhibits fast mixing as well as a considerable reduction in computational cost per iteration. PICAR is computationally efficient and scales well to high dimensions. It is also automated and easy to implement for a wide array of user-specified hierarchical spatial models. We show, via simulation studies, that our approach performs well in terms of parameter inference and prediction. We provide several examples to illustrate the applicability of our method, including (i) a high-dimensional cloud cover dataset that showcases its computational efficiency, (ii) a spatially varying coefficient model that demonstrates the ease of implementation of PICAR in the probabilistic programming languages stan and nimble, and (iii) a watershed survey example that illustrates how PICAR applies to models that are not amenable to efficient inference via existing methods.

Sequential Design of Multi-Fidelity Computer Experiments: Maximizing the Rate of Stepwise Uncertainty Reduction

P. 199-209

Rémi Stroh, Julien Bect, Séverine Demeyer, Nicolas Fischer, Damien Marquis, Emmanuel Vazquez

Abstract

This article deals with the sequential design of experiments for (deterministic or stochastic) multi-fidelity numerical simulators, that is, simulators that offer control over the accuracy of simulation of the physical phenomenon or system under study. Accurate simulations usually entail a high computational effort, while coarse simulations are obtained at a lower cost. The cost can be measured, for example, by the run time of the simulator or the financial cost of the computing resources. In this setting, simulation results obtained at several levels of fidelity can be combined in order to estimate quantities of interest (the optimal value of the output, the probability that the output exceeds a given threshold, etc.) in an efficient manner. We propose a new Bayesian sequential strategy called maximal rate of stepwise uncertainty reduction (MR-SUR), that selects additional simulations to be performed by maximizing the ratio between the expected reduction of uncertainty and the cost of simulation. This generic strategy unifies several existing methods, and provides a principled approach to develop new ones. We assess its performance on several examples, including a computationally intensive problem of fire safety analysis where the quantity of interest is the probability of exceeding a tenability threshold during a building fire.

Monitoring Heterogeneous Multivariate Profiles Based on Heterogeneous Graphical Model

P. 210-223

Hui Wu, Chen Zhang, Yan-Fu Li

Abstract

Process monitoring using profile data remains an important and challenging problem in various manufacturing industries. Motivated by an application case of motherboard testing processes, we develop a novel modeling and monitoring framework for heterogeneous multivariate profiles. In this framework, a heterogeneous graphical model is constructed to depict the complicated heterogeneous relationship among profile channels. Then monitoring the heterogeneous relationship among profile channels can be reduced to monitoring the graphical networks. Besides, we investigate several theoretical results concerning the accuracy of the estimated graphical structure. Finally, we demonstrate the proposed method through extensive simulations and a real case study.

Bayesian Dynamic Feature Partitioning in High-Dimensional Regression With Big Data

P. 224-240

Rene Gutierrez & Rajarshi Guhaniyogi

Abstract

Bayesian computation of high-dimensional linear regression models using Markov chain Monte Carlo (MCMC) or its variants can be extremely slow or completely prohibitive since these methods perform costly computations at each iteration of the sampling chain. Furthermore, this computational cost cannot usually be efficiently divided across a parallel architecture. These problems are aggravated if the data size is large or data arrive sequentially over time (streaming or online settings). This article proposes a novel dynamic feature partitioned regression (DFP) for efficient online inference for high-dimensional linear regressions with large or streaming data. DFP constructs a *pseudo posterior density* of the parameters at every time point, and quickly updates the pseudo posterior when a new block of data (data shard) arrives. DFP updates the pseudo posterior at every time point suitably and partitions the set of parameters to exploit parallelization for efficient posterior computation. The proposed approach is applied to high-dimensional linear regression models with Gaussian scale mixture priors and spike-and-slab priors on large parameter spaces, along with large data, and is found to yield state-of-the-art inferential performance. The algorithm enjoys theoretical support with pseudoposterior densities over time being arbitrarily close to the full posterior as the data size grows, as shown in the supplementary material. Supplementary material also contains details of the DFP algorithm applied to different priors. Package to implement DFP is available in https://github.com/Rene-Gutierrez/DynParRegReg. The dataset is available in https://github.com/Rene-Gutierrez/DynParRegReg_Implementation.

Anomaly Detection in Large-Scale Networks With Latent Space Models

P. 241-252

Wesley Lee, Tyler H. McCormick, Joshua Neil, Cole Sodja, Yanran Cui

Abstract

We develop a real-time anomaly detection method for directed activity on large, sparse networks. We model the propensity for future activity using a dynamic logistic model with interaction terms for sender- and receiver-specific latent factors in addition to sender- and receiver-specific popularity scores; deviations from this underlying model constitute potential anomalies. Latent nodal attributes are estimated via a variational Bayesian approach and may change over time, representing natural shifts in network activity. Estimation is augmented with a case-control approximation to take advantage of the sparsity of the network and reduces computational complexity from $O(N2)$ to $O(E)$, where N is the number of nodes and E is the number of observed edges. We run our algorithm on network event records collected from an enterprise network of over 25,000 computers and are able to identify a red team attack with half the detection rate required of the model without latent interaction terms.

An Adaptive Sampling Strategy for Online Monitoring and Diagnosis of High-Dimensional Streaming Data

P. 253-269

Ana María Estrada Gómez, Dan Li, Kamran Paynabar

Abstract

Statistical process control techniques have been widely used for online process monitoring and diagnosis of streaming data in various applications, including manufacturing, healthcare, and environmental engineering. In some applications, the sensing system that collects online data can only provide partial information from the process due to resource constraints. In such cases, an adaptive sampling strategy is needed to decide where to collect data while maximizing the change detection capability. This article proposes an adaptive sampling strategy for online monitoring and diagnosis with partially observed data. The proposed methodology integrates two novel ideas (i) the recursive projection of the high-dimensional streaming data onto a low-dimensional subspace to capture the spatio-temporal structure of the data while performing missing data imputation; and (ii) the development of an adaptive sampling scheme, balancing exploration and exploitation, to decide where to collect data at each acquisition time. Through simulations and two case studies, the proposed framework's performance is evaluated and compared with benchmark methods.

Technometrics, ISSN 0040-1706
Volume 64, number 2 (may 2022)