Proposta di assegno di ricerca

Project title:

Projection-based depth function and its use in classification and outlier detection

Supervisor of the project:

Prof. Cinzia Viroli

Brief introduction:

The purpose of this project is to introduce and explore a new concept of projection-based depth function as a powerful tool for classification and outlier detection. By utilizing this innovative approach, we aim to enhance the accuracy and robustness of classification models, as well as effectively identify and handle outliers in various data sets. We will investigate the theoretical foundations of projection-based depth function and its practical implications in real-world applications. Through this research, we seek to advance the field of data analysis by providing new insights into classification techniques and outlier detection methods, ultimately contributing to more accurate and reliable data-driven decision-making processes.

Background and statement of the problem:

In multivariate analysis, the task of identifying order statistics, quantiles, and typical or atypical patterns poses significant challenges due to the absence of a natural ordering among observations, unlike the case in the real line (Kong & Mizera, 2012; Serfling, 2002).

To address this issue, a prominent line of research has emerged based on the concept of statistical depth, which establishes a natural center-outward ordering of sample points in R^p (where $p>1$). A depth function assigns a real number to each point in a multivariate dataset, measuring the outlyingness of the point relative to the dataset's barycenter. It quantifies the distance of an observation from the center of the dataset and is utilized to identify order statistics using depth-induced contours.

Several popular depth functions have been developed. For instance, the halfspace depth (Tukey, 1975) measures the depth of a point as the probability that it lies inside a randomly selected halfspace containing the dataset's center. The Mahalanobis depth (Liu, 1993) is another widely used depth function. Additional depth functions, such as simplicial depth, regression depth, and majority depth, have been introduced and applied in diverse contexts, including quality control in manufacturing and exploratory statistical analysis. Liu (1999) and Serfling (2000) provided different depth functions and formulated a comprehensive and constructive definition based on desirable properties.

In this research, we start by a novel definition of a depth function for multivariate data using random spherical directions. This method is designed to satisfy the essential properties of depth functions (Serfling, 2000, 2002) when the data is sphered. Furthermore, it preserves the Mahalanobis distance of the multivariate points in the original p-dimensional space. Importantly, this definition fully characterizes the probability distribution of the data. Specifically, the proposed depth function is the expectation of all depths along potentially infinite random directions, which themselves depend on point percentiles estimated using parametric or nonparametric models. Common choices are given by the Gaussian and the kernel models. In addition, the flexibility of the fgld quantile distribution (Redivo et al., 2023; Chakrabarty et al., 2021) offers valuable options within the model choices in the projected spaces.

Given that the proposed directional distribution depth function allows for a population version, it can be employed for out-of-sample prediction. This enables its utilization as a tool in supervised classification problems, where it provides a distance measure for test set points relative to the distributions of different classes.

The goal of this research proposal is to extend the method to unsupervised classification and outlier detection. The performance of the proposed methods will be evaluated through simulated experiments and real data applications, in comparison to alternative classification and outlier detection methods.

Research hypothesis, aim, objectives and deliveries:

We aim to address four key points encompassing methodological and applied developments in the context of projection-based depth function and its use in classification and outlier detection. The following aspects outline our research objectives:

1. Unsupervised Classification: Our first objective is to conduct a comprehensive clustering analysis using the projection-based depth function. This analysis will provide insights into the underlying structure of the data and identify distinct clusters or groups. Initially, we will perform exploratory analyses, such as descriptive statistics and visualization techniques, to gain an understanding of the data's characteristics. Subsequently, we will employ the projection-based depth function to establish a natural ordering of data points and facilitate clustering. We will evaluate the performance of the proposed approach using benchmark datasets and compare it with existing clustering techniques. The R programming language, along with suitable packages, will be utilized for implementing the clustering analysis.

2. Outlier Detection: The second objective of this research is to develop an effective outlier detection methodology based on the projection-based depth function. Outliers are data points that deviate significantly from the majority of observations and can provide valuable insights or indicate potential errors or anomalies in the data. By leveraging the projection-based depth function, we will quantify the outlyingness of individual data points relative to the dataset's center and structure. We will investigate various thresholding techniques and statistical metrics to determine the presence of outliers based on the depth values. The performance of our approach will be evaluated using synthetic datasets with known outliers and real-world datasets, comparing it with established outlier detection methods. The implementation of this methodology will be carried out using the R programming language, incorporating relevant packages and libraries.

3. Theoretical and Asymptotic Properties: The third objective of this research is to analyze the theoretical and asymptotic properties of the proposed projection-based depth function method. We will explore the statistical properties of the method, including consistency, efficiency, and robustness. Theoretical derivations and proofs will be conducted to establish the theoretical foundations of the methodology. Furthermore, we will investigate the asymptotic behavior of the method under different assumptions and sample size scenarios. These analyses will contribute to understanding the statistical properties and limitations of the projection-based depth function method and provide insights into its applicability in various contexts.

4. Simulation and Real Data Analysis: The final step of this research is to assess the performance and robustness of the projection-based depth function in classification and outlier detection through simulation studies and real data analysis. We will conduct extensive simulations to evaluate the behavior and effectiveness of the proposed methodology under various data scenarios, including different cluster structures, varying degrees of outliers, and varying data dimensions. Additionally, we will apply the developed methodology to real-world datasets relevant to the classification and outlier detection tasks to examine its applicability and performance in practical settings. The results of the simulation studies and real data analysis will provide valuable insights into the strengths and limitations of the projection-based depth function for classification and outlier detection tasks.

The post-doctoral researcher assigned to this project will be responsible for gaining a deep understanding of the projection-based depth function and its applications in classification and outlier detection. They will research and propose innovative solutions aligned with the research objectives outlined above and implement them in R. Additionally, the researcher will perform extensive simulation studies and analyze real-world datasets to evaluate the performance and practicality of the proposed methodology. The results of the research will be presented at conferences and submitted for publication in scientific statistical journals. Collaboration with international partners will be sought to foster knowledge exchange and enhance the research outcomes.

References

Chakrabarty, T. K. and Sharma, D. (2021). A Generalization of the Quantile-Based Flattened Logistic Distribution. Annals of Data Science, 8(3): 603-627.

Kong, L. and Mizera, I. (2012). Quantile tomography: Using quantiles with multivariate data. Statistica Sinica, 22(4): 1589-1610.

Kong, L. and Zuo, Y. (2010). Smooth depth contours characterize the underlying distribution. J. Multivariate Analysis, 101(9): 2222-2226.

Koshevoy, G. and Mosler, K. (1997). Zonoid trimming for multivariate distributions. Ann. Stat., 25(5):1998-2017.

Koshevoy, G. A. (2002). The Tukey Depth Characterizes the Atomic Measure. J. Multivariate Anal., 83(2): 360-364.

Liu, R., Parelius, J., and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. Annals of Statistics, 27(3): 783-858.

Mahalanobis, P. (1936). On the generalized distance in statistics. Proceedings of the National Institute of Science of India, 2: 49-55.

Nagy, S. (2021). Halfspace depth does not characterize probability distributions. Statist. Papers, 62(3): 1135-1139.

Redivo, E., Viroli, C., and Farcomeni, A. (2023). Quantile-based distribution functions and their use for classification, with application to naive bayes classifiers. Statistics and Computing, forthcoming.

Sering, R. (2002). Quantile functions for multivariate analysis: approaches and applications. Statistica Neerlandica, 56(2):214-232.

Sering, R. and Zuo, Y. (2000). General notions of statistical depth function. Ann. Stat., 28(2): 461-482.

Struyf, A. J. and Rousseeuw, P. J. (1999). Halfspace Depth and Regression Depth Characterize the Empirical Distribution. J. Multivariate Anal., 69(1): 135-153.

Zuo, Y. and Sering, R. (2000). General notions of statistical depth function. The Annals of Statistics, 28(2): 461-482.