Categorical Exploratory Data Analysis

Subjects: Computer Science, Artificial Intelligence | Statistics & Probability
Contributor: Elizabeth Chou

Categorical exploratory data analysis (CEDA) is demonstrated to provide new resolutions for two topics: multiclass classification (MCC) with one single categorical response variable and response manifold analytics (RMA) with multiple response variables.

Keywords: multiclass classification ; categorical exploratory data analysis ; PITCHf/x

1. Overview

All features of any data type are universally equipped with categorical nature revealed through histograms. A contingency table framed by two histograms affords directional and mutual associations based on rescaled conditional Shannon entropies for any feature-pair. The heatmap of the mutual association matrix of all features becomes a roadmap showing which features are highly associative with which features. We develop our data analysis paradigm called categorical exploratory data analysis (CEDA) with this heatmap as a foundation. CEDA is demonstrated to provide new resolutions for two topics: multiclass classification (MCC) with one single categorical response variable and response manifold analytics (RMA) with multiple response variables. We compute visible and explainable information contents with multiscale and heterogeneous deterministic and stochastic structures in both topics. MCC involves all feature-group specific mixing geometries of labeled high-dimensional point-clouds. Upon each identified feature-group, we devise an indirect distance measure, a robust label embedding tree (LET), and a series of tree-based binary competitions to discover and present asymmetric mixing geometries. Then, a chain of complementary feature-groups offers a collection of mixing geometric pattern-categories with multiple perspective views. RMA studies a system's regulating principles via multiple dimensional manifolds jointly constituted by targeted multiple response features and selected major covariate features. This manifold is marked with categorical localities reflecting major effects. Diverse minor effects are checked and identified across all localities for heterogeneity. Both MCC and RMA information contents are computed for data's information content with predictive inferences as by-products. We illustrate CEDA developments via Iris data and demonstrate its applications on data taken from the PITCHf/x database.

2. Data Analysis

The author of the well-known 1977 book Exploratory Data Analysis (EDA) [1], John W. Tukey in his 1962 paper [2] The future of Data Analysis discussed fundamental principles and made clear-cut arguments for data analysis as a scientific discipline. Among many facts, he emphasized that data analysts must put "reliance upon the test of experience as the ultimate standard of validity". He further wrote that: "Data analysis is a larger and more varied field than inference, or incisive procedures, or allocation."

This statement rings much louder now than ever before in this big data era. When a system of interest is somehow complex and being embraced by a big dataset, the goal of data analysis is naturally data's full information content for system understanding. Many tasks can be performed as by-products based on such information content.

Within the 60 years from 1962, the data visualization technique is the most visible kind of effort developed for carrying out EDA in the literature [3][4][5]. Some attempts sparsely appeared in the literature that intend to incorporate EDA into the Bayesian framework [6]. Thus far, we have seen neither well-developed resolutions for major problems listed in Tukey's paper, including the multiple response problem, nor unified fundamental concepts and computational paradigms as principles of data analysis. Although data analysis nowadays has been widely permeating and drastically expanding in all sciences, one factual sign is continuously blinking: data analysis is still underdeveloped [5][7].

One direct and clear view of such underdevelopment is seen when analyzing a real database from a somehow known complex physical system. Physics tells us the principles that are underlying the system and how they work. However, we often lack knowledge of how these principles realistically couple and link together in working out details. The phrase "The

devil is in the detail" seems to capture what is precisely missing in data analysis. Can data analysts discover such principles and missing details of their linkages? Would our data analysis offer a complete understanding of a system of interest? We attempt positive answers to both questions through real complex systems in this paper.

The major real-world complex system considered here is the Major League Baseball's (MLB) pitching dynamics. In the USA, MLB has been recording and storing every single pitch in its 30 stadiums into its public available PITCHf/x database since 2006. This is one of the best-maintained databases in the world and one very important database for the evolutions of thousands of pitchers' aerodynamics and biomechanics of pitching. The importance of this database is far beyond baseball as a sport in this big data era since it adds about 700k pitches, each of which is more than 30D data, every season. In this paper, we focus on three complex systems defined by three sets of selected pitchers' three pitching types.

From the physics perspective, these complex systems involve Newtonian laws of forces, as well as Magnus force of spin [8]. Both forces govern a baseball trajectory. We want to compute details of pitching dynamics contained in the PITCHf/x database without invoking differential equations from physics and aerodynamics literature. Such computable details would make us see which factors make a pitch move and curve the way it does and understand pitchers' idiosyncratic characteristics. Here, well-patched details mean multiscale and global-to-local pattern information illuminating how the principles work in concert. We collectively term such details as data's "information content".

Here, PITCHf/x's information content regarding pitching dynamics contains at least two major perspectives: (1) what factors can efficiently characterize a pitcher's pitches; and (2) how the underlying physical principles work. The first perspective is one major classification topic, called multiclass classification (MCC) in machine learning (ML). The second one is the topic called response manifold analytics (RMA) for solving multiple response problems. Here, a manifold would naturally appear for depicting how a group of possibly highly associative response features or variables globally links to a major group of covariate features. Upon this manifold, other minor groups of covariate features can also involve locally, not globally. This manifold-based topic is not yet well developed in the statistics and ML literature.

Further in PITCHf/x example, the response and covariate features have a clear spatial-temporal separation. Covariate features are measured at the pitcher's mound, while response features are measured near the home plate. Thus, they are not exchangeable. Although RMA is exactly one physical multiple response problem here, its manifold framework is seemingly universal when continuous features of physical or mechanical mechanisms are intertwined and linked. We expect that RMA's information content should bear with principles on global and large scales and reveals heterogeneous effects on local and fine scales. Data's deterministic and stochastic structures are all presented through visible and explainable graphic displays based on such information content. These computed structures, in turn, serve as bases for our inferential decision-making. In this fashion, our data analysis indeed coherently reflects the quoted statement at the beginning of this section.

As the first phase of our CEDA developments, we consider only quantitative covariate features under both settings of MCC and RMA here. As such, the locality concept is either based on hypercubes or *k*-nearest neighbors (KNN). In a separate report, the second phase of CEDA is developed to unify both topics simultaneously and accommodate categorical covariate features. Throughout these two phases of CEDA developments, the concepts and devices of histogram, contingency table, and mutual conditional entropy (MCE) matrix constantly play fundamental and critical roles.

3. Conclusions

Under the setting of quantitative covariate features, we develop computational algorithms and protocols of CEDA for MCC and RMA to extract data's multiscale information contents with heterogeneity. Such resultant information contents echo physicist P. W. Anderson's universal view on large complex systems. By fully reflecting the geometries and manifolds underlying complex systems of interest, we demonstrate that the chief merits of our computational developments are geared for system understanding. Upon three relatively small datasets taken from the PITCHf/x database, we also point out that inferential decision-making based on data's information content is likely universally valid. Hopefully, this conclusive message can bear essential impacts on any other system studied far beyond MLB's pitching dynamic systems.

On the front of CEDA for MCC, our developments of a chain of complementary feature-groups and tabular representations of the resultant collection of mixing geometric pattern-categories explicitly reveal asymmetry of serial mixing geometries as MCC's information content. Such discoveries and synthesis reiterate the essence of CEDA for MCC in discovering diverse perspectives of heterogeneity within any dataset from a complex system. Our partial ordering and dominance matrix concepts for computing a label embedding tree (LET) are robust and efficient. Consequently, With LET's reliable binary geometry on a label space, our serial binary competitions with length of order $O(\log_2 K)$ are an effective way of

exploring complex mixing geometry among many high dimensional point-clouds. It is worth mentioning that it is to the great benefit of ML's MCC results by coupling with results of CEDA for MCC because its certainty and uncertainty are documented and explained.

On the front of CEDA for RMA, we show that manifolds based on response-to-covariate associations in Euclidean spaces are efficient foundations for studying complex systems. Such manifolds pave the platform for figuring out data's information content with locality-based heterogeneity and facilitate fundamentally efficient and straightforward inferential protocols, which are particularly suited for predictive decision-making on multiple response variables.

At the end of this section, we remark that the second phase of CEDA development is undergoing research in a separate study. This phase of CEDA would have the capacity of accommodating categorical covariate features and other complex settings. After the second phase of development, CEDA can become one of the foundations of data analysis on a structured data matrix. As such, CEDA would be a standard protocol for studying complex physical systems. That is, data analysis might realistically become a scientific discipline on its own, as John Tukey projected in 1962.

References

1. Tukey, J.W. Exploratory Data Analysis; Pearson: London, UK, 1977.
2. Tukey, J.W. The Future of Data Analysis. Ann. Math. Stat. 1962, 33, 1–67.
3. Tufte, E.R. The Visual Display of Quantitative Information; Graphics Press: Cheshire, CT, USA, 1983.
4. Wilkinson, L. The Grammar of Graphics, 2nd ed.; Springer: New York, NY, USA, 2005.
5. Gelman, A.; Vehtari, A. What are the most important statistical ideas of the past 50 years? arXiv 2020, arXiv:2012.00174.
6. Gelman, A. A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. Int. Stat. Rev. 2003, 71, 369–382.
7. Donoho, D.L. 50 years of data science. J. Comput. Graph. Stat. 2017, 26, 745–766.
8. Briggs, L. Effect of Spin and Speed on the Lateral Deflection (Curve) of a Baseball; and the Magnus Effect for Smooth Spheres. Am. J. Phys. 1959, 27, 589–596.