Radar-boxplot: a boxplot variation for multivariate classification data

Author 1*
Department of YYY, University of XXX
and
Author 2
Department of ZZZ, University of WWW

November 29, 2018

Abstract

The text of your abstract. 200 or fewer words.

Keywords: 3 to 6 keywords, that do not appear in the title

*The authors gratefully acknowledge *please remember to list all relevant funding sources in the unblinded version*

1 Introduction

As we increase our potential to collect and analyze data, the datasets become larger, more complex and harder to understand. Visualizing classification multivariate data has been a non trivial task, relying on methods to reduce data dimensionality which may over simplify the problem and fail to represent the whole complexity.

In the realm of classification problem, there has been great advances, mainly due to increase in computational power, allowing to use complex machine learning algorithms. These algorithms have provided means for interpreting and classifying data which were otherwise meaningless or could only be understood by the human eye. But those algorithms don't solve the problem of understanding how and why the data is meaningful, relegating the comprehension of the problem to the background.

Although practical applications don't really require to fully understand how good results are obtained, for scientific applications that is important to test and purpose solutions based on evidences rather than reach the solution by trial and error.

Traditionally pattern recognition multivariate data can be visualized by two means: 1. Using multiple boxplots for each variable, discriminating the classes by colors (Pu2012). 2. Plotting a 2D or at most 3D scatterplot reducing the variables by selecting most important variables or transforming them to new vectors of data using methods such as Principal Component Analysis (PCA) and Linear Discriminant Analysis.

When using multiple boxplots, although it is possible to compare classes for each variable, one loses track of the overall multivariate pattern showed by a class.

Plotting 2D or 3D scatterplots requires to reduce the problem to at most three variables. Whereas the attribute selection may leave out important variables which would characterize the different classes, reducing dimensionality through PCA or such methods can lead to good separability, but the correspondence from the original variables to the derived components is not straightforward, so it is not possible to really understand how the original variables defines the classes pattern.

Exemplos de como os dados são analisados hoje em dia.

2 Methods

The radar-boxplot is plotted using the radar plot design, along with the traditional boxplot concepts proposed by FISHER: procurar.

While the radar plot can display multivariate data, it fails to represent groups, instead it plots individual features, whereas boxplot represent the distribution of a group but for a single variable. Merging the two concepts, we can visualize the distribution of a group for multiple variables in a single plot.

The concept is simple. We create a polar plot, dividing it by the number of variables we need to represent, each axis is a variable. Then, it is possible to represent the Q25-Q75 as a single radial polygon, with a varying width along each axis, whereas the Q0-Q25 and Q75-Q100 will be represented by two other radial polygons along the margins of the first Q25-Q75 polygon. The outliers as defined in the boxplot and the median can also be represented as dots and a circular line.

The polar plot is drawn as ranging from 0 to 1, from the center out. The variables must be standardized to range from 0.1 to 1 to be drawn in the same scale. It is important that the minimum value is not 0, otherwise it will be difficult to visualize low values as they will overlap each other.

3 Results

Gráficos

1. Geral, radar-boxplot com multiplas classes: mostrando como o grafico consegue mostrar o panorama geral e revelar padroes que se confirmam pela acuracia atingida com os classificadores

2. Matriz de erros para fazer o paralelo com as classes: Fazendo o paralelo com o grafico geral, demonstrando a correspondencia entre os padroes revelados pelo grafico e a confusao encontrada pelo classificador

3. Gráfico radar-boxplot com feicoes classificadas erradas: demonstrando a possibilidade de verificar caso a caso se o individuo segue ou nao o padrao, apontando possibilidades

para melhorar a classificacao

4 Discussion

The radar-boxplot provides an intuitive way for visualizing multivariate classification data.