Principal Component Analysis of data complexity measures for cancer breast prediction

Mario Cobos, Armando Collado-Villaverde, Daniel Rodriguez and David F. Barrero

Universidad de Alcalá, Madrid, Spain

mario.cobos@edu.uah.es, {armando.collado, daniel.rodriguezg, david.fernandezb}@uah.es

UNIÓN EUROPEA
Fondo Social Europeo
Iniciativa de Empleo Juvenil
El FSE invierte en tu futuro

Comunidad de Madrid

Introduction

- Machine Learning algorithms require large quantities of data and time
- Data complexity can be characterized using the twelve metrics defined by Ho and Basu [TB02]
- We confirm the results obtained by the original authors, and attempt to correlate data complexity and classifier quality using a well-known dataset
  - In order to do so, we use k-fold cross validation repeatedly
  - To correct potential outliers, we repeat the experiment a number of times for each value of k

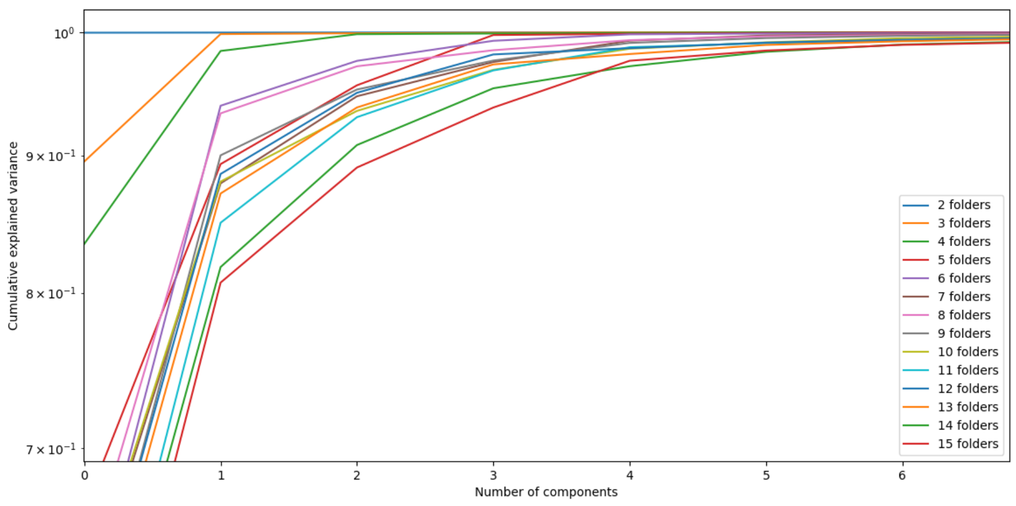
Complexity measures

- Measures of overlaps:
  - F1: Maximum Fisher's discriminant ratio
  - F2: Volume of overlap region
  - F3: Maximum feature efficiency
- Measures of class separability:
  - L1: Minimized sum of error distance by linear programming
  - L2: Error rate of linear classifier by LP
  - N1: Fraction of points on class boundary
  - N2: Ratio of average intra/inter class NN distance
  - N3: Error rate of 1NN classifier
- Measures of geometry, topology and density of manifolds:
  - L3: Nonlinearity of a linear classifier by LP
  - N4: Nonlinearity of 1NN classifier
  - T1: Fraction of points with associated adherence subsets retained
  - T2: Average number of points per dimension

Principal Component Analysis

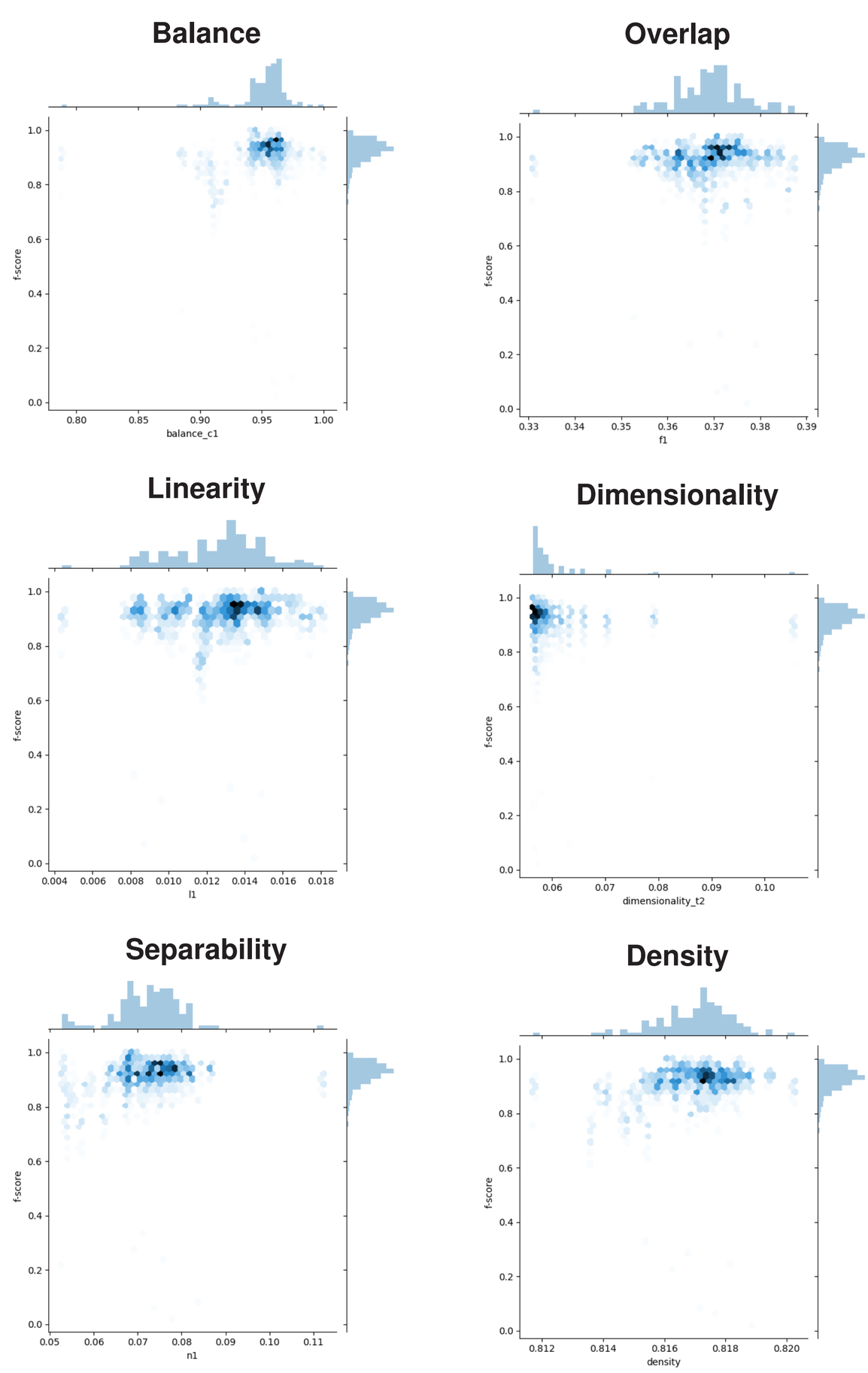
- Inputs:
  - Data complexity measures obtained from the different folders
- Outputs:
  - Number of metrics required to determine the quality of the data

Number of components required to explain the variance according to the number of folders



- A total of six components is required when k reaches its peak value of fifteen
- The smaller the dataset, the more components are required in order to explain the variance obtained
- Most of the variance can be explained using only the first component, independently of the number of folders
- Increasing the number of sample for a given value of k does not meaningfully change the form of the plot presented above
- The results obtained align with those described by Ho and Basu [TB02]

Correlation between complexity and quality



- The dataset selected provides highly balanced, high quality data
  - The resulting classifiers tend to be accurate as a result
  - The range of the metrics obtained displays a meaningful relationship between the quality of the dataset and the quality of the resulting classifier

Conclusions

- The quality of a dataset has an impact on the quality of the resulting classifier
- The metrics defined to measure dataset quality can be reduced to a relatively small number of components

Acknowledgements

The work is co-funded by the European Social Fund, Comunidad de Madrid *Garantía Juvenil* (PEJD-2018-PRE/TIC-8176) and Junta de Comunidades de Castilla-La Mancha (SBPLY/18/180501/000019)

References

Tin Kam Ho and M. Basu.
Complexity measures of supervised classification problems.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3):289–300, March 2002.