Supervised classification of emission stars spectra

Jaroslav Vážný, Petr Škoda

Astronomical Institute of the Academy of Sciences of the Czech Republic

Abstract

Advances in the technology of astronomical spectra acquisition have resulted in an enormous amount of data available in world-wide telescope archives. It is no longer feasible to analyze them using classical approaches, so a new astronomical discipline, astroinformatics, has emerged. We describe the initial experiments in the investigation of spectral line profiles of emission line stars using machine learning with attempt to automatically identify Be and B[e] stars spectra in large archives and classify their types in an automatic manner. Due to the size of spectra collections, the dimension reduction techniques based on wavelet transformation are studied as well. The result clearly justifies that machine learning is able to distinguish different shapes of line profiles even after drastic dimension reduction.

Supervised Classification of Emission Stars Spectra

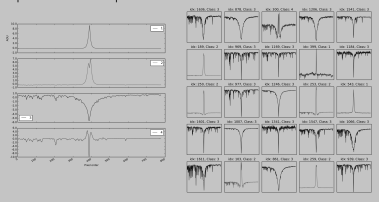
Jaroslav Vážný, Petr Škoda
Astronomical Institute of the Academy of Sciences of the Czech Republic

Objectives

1. Classify emission star spectra based on H-α line profile.
2. Compare methods for dimensionality reduction.
3. Tune classifier parameters.

Introduction

- There are 1805 manually classified spectra from Ondřejov observatory divided into 4 categories based on profile of the H-α line. We want to train a SVM classifier to automatically categorize the rest. Figure on the left shows the average spectrum in each category and on the right there are samples of individual spetra.

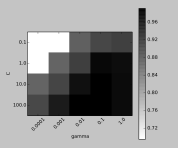


Dimensionality Reduction

- Each spectrum has 778 points. Different approaches were tested for dimensionlity reduction.
  - PCA: Principal component analysis. Linear dimensionality reduction using Singular Value Decomposition of the data and keeping only the most significant singular vectors to project the data to a lower dimensional space.
  - Isomap: Isometric Mapping. Can be viewed as an extension of Multi-dimensional Scaling (MDS) or Kernel PCA. Isomap seeks a lower-dimensional embedding which maintains geodesic distances between all points.
  - LLE: Locally linear embedding seeks a lower-dimensional projection of the data which preserves distances within local neighborhoods. It can be thought of as a series of local Principal Component Analyses which are globally compared to find the best non-linear embedding. [1]

Parameters tuning

- Support Vector Machines (SVM) was used as a classifier with Radial Basis Function (RBF) kernal ($\exp(-\gamma|\mathbf{x}-\mathbf{x}'|^2)$). Grid search was used to find optimal values for $\mathbf{C}$ and γ parameters. There is an example of heatmap for optimal parameters for data reduced by Isomap method.



Contact Information

- Web: `http://physics.muni.cz/~vazny/wiki/index.php`
- Email: jaroslav.vazny@gmail.com
- Phone: +42 606 77 65 64

Acknowledgment

- This research was supported by grant GAČR 13-08195S.

Classification

- Data were splited randomly into training (75%) and testing (25%) sample. Reports for different reduction aproaches are shown below compared to non-reduced data sample. Precision = tp / (tp + fp), recall = tp / (tp + fn), f1-score = 2 * (precision * recall) / (precision + recall) where tp is true positive, fp false negative and fn false negative.

Category	Precision	Recall	f1-score	Support
1	0.95	0.95	0.95	61
2	0.96	0.96	0.96	74
3	1.00	1.00	1.00	303
4	0.93	0.93	0.93	14
avg/total	0.98	0.98	0.98	452

Table 1: Classification report without dimensionality reduction

Category	Precision	Recall	f1-score	Support
1	0.89	0.95	0.92	61
2	0.96	0.91	0.93	74
3	1.00	1.00	1.00	303
4	1.00	0.93	0.96	14
avg/total	0.98	0.98	0.98	452

Table 2: ISOMAP Classification report

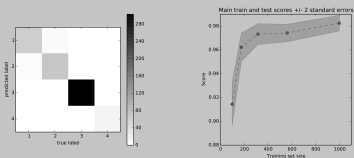
Category	Precision	Recall	f1-score	Support
1	0.97	1.00	0.98	61
2	1.00	0.97	0.99	74
3	0.99	1.00	1.00	303
4	1.00	0.86	0.92	14
avg/total	0.99	0.99	0.99	452

Table 3: PCA Classification report

Category	Precision	Recall	f1-score	Support
1	0.91	1.00	0.95	61
2	0.99	0.92	0.95	74
3	1.00	1.00	1.00	303
4	0.92	0.86	0.89	14
avg/total	0.98	0.98	0.98	452

Table 4: LLE Classification report

The left figure shows classification confusion matrix. Graph on the right is a learning curve.



Conclusion

- It is possible to dramatically reduce the number of dimensions of spectra in classification problem (here from 778 to 10).
- PCA, Isomap, LLE and has similar effects in this classification problem.
- It is always important to tune corresponding hyperparameters.

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.
Scikit-learn: Machine learning in Python.
Journal of Machine Learning Research, 12:2825–2830, 2011.

[2] Scott F Daniel, Andrew Connolly, Jeff Schneider, Jake Vanderplas, and Liang Xiong.
Classification of stellar spectra with local linear embedding.
The Astronomical Journal, 142(6):203, 2011.