university of groningen / faculty of science and engineering

Classification and Feature Relevance Determination for Imbalanced Astronomical Data

Jarvin Mutatiina

SUPERVISORS
Asst. Prof. Dr Kerstin Bunte
Prof. Michael Biehl
Mohammad Mohammadi, MSc
Teymoor Saifollahi, MSc

Introduction

Introduction

- Data Science in Astronomy is becoming more popular with increasingly complex and vast amounts of data available.
- Common conventional astronomical method of analysis is ***Photometric selection***[Stetson, 2013]
- ML techniques have been previously explored e.g. SVM, KNN,Random Forest, Artificial Neural Networks etc.

Introduction

Need to explore more sophisticated and robust methods of *adaptive distance metric* variants of ***Learning Vector Quantization***(LVQ).

Why?

- Distance metric is adapted to the target data as opposed to commonly used Euclidean distance.
- Very intuitive and highly interpretable with the output feature relevance.
- Scalable because of the prototype class representation.

Research contribution/Objective

- Application of state of the art adaptive distance metric LVQ variants to imbalanced astronomical data.
- Handling of class imbalance and overlap with pre-processing techniques
- Feature relevance determination and relation to astronomical expert knowledge.

Data Definition

Data Definition

Data is extracted from images from different points of observations through optical filters u, g, r, i, J, K using a Source Extractor (SExtractor)[Bertin and Arnouts, 1996].
There are 3 classes;

- **`Class 1`** - Background galaxies
- **`Class 2`** - Ultra-compact galaxies(UCDs) and Globular clusters (GCs). Minority class of interest.
- **`Class 3`** - Foreground stars.

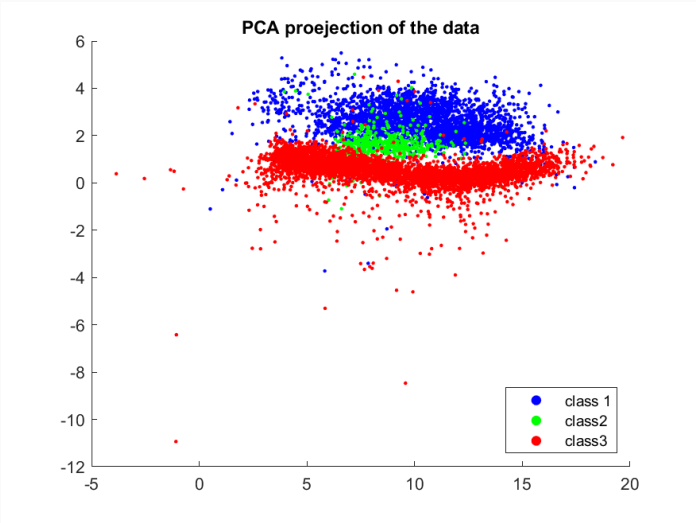
Data Definition

There are 21 features;

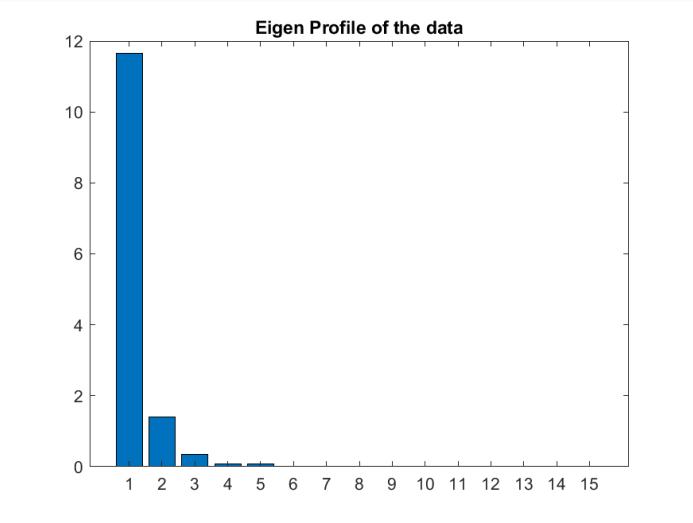
- Proxies of the Angular size i.e Full-width-half-maximum - $fwhm^*g, fwhm^*r, fwhm^*i, fwhm^*u, fwhm^*j, fwhm^*k$
- Compactness of the objects - $dg0, di0, di0$
- Color indices - $ug, ur, ui, uj, uk, gr, gi, gj, gk, ri, rj, rk, ij, ik, jk$

	Class 1	Class 2	Class 3	Total
data1	2003	163	4121	6287
data2	2826	512	4399	7737
	2083	387	2806	5276

Data Structure/ Projection



(a)



(b)

Figure 1: (a) PCA-projection of the *data2*. It shows the imbalance and overlap between the classes. (b) The eigen profile gives a hint that there are 2-3 more significant directions of the data.

Photometric Selection

Is traditionally used by astronomers on basis of a color-color diagram. The color indices **ui**- age of the object and **ik** - metallicity are most instrumental according to Munoz et. al[Munoz et al., 2013].

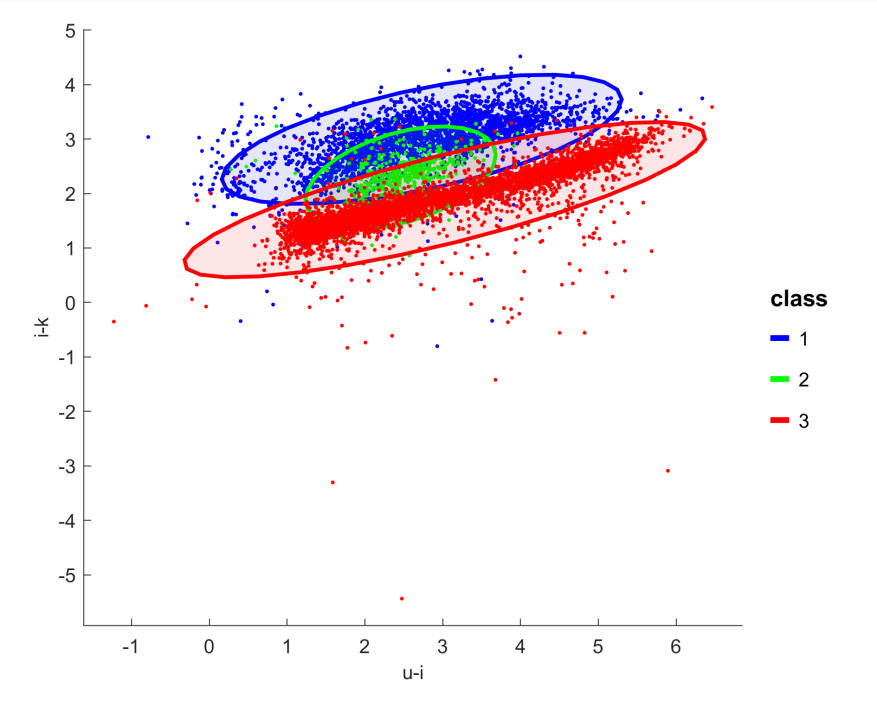


Figure 2: Color-color diagram showing photometry-based selection with the **ui** and **ik** colors. Based on the position of novel data points, a class can be assigned according to its position

Methods

Data Pre-processing

Appropriate pre-processing is crucial for this research

- Z-score Transformation
- Synthetic Minority Over-Sampling (SMOTE)[Chawla et al., 2002]

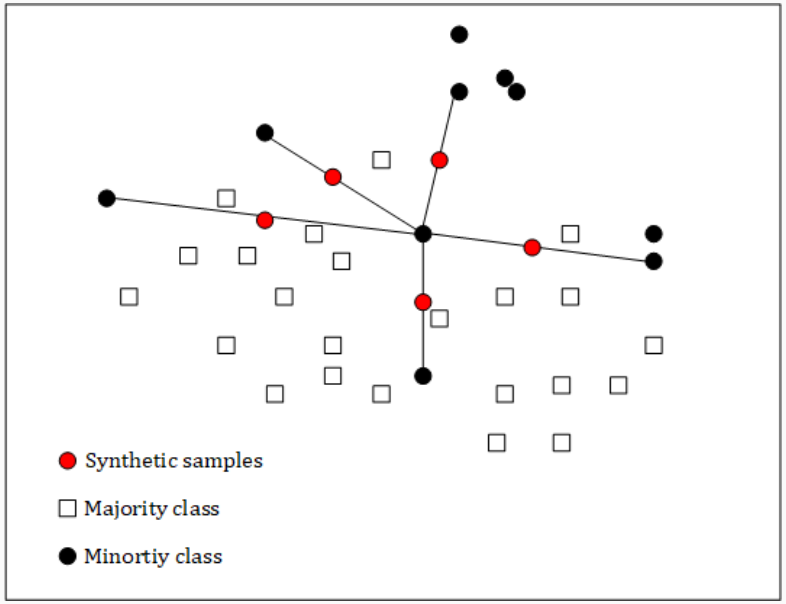


Figure 3

- Borderline-SMOTE[Han et al., 2005]

Learning Vector Quantization -LVQ

- Is a prototype based ML method based on similarity measure introduced by Kohonen[Kohonen, 1997].
  For data ξ_i of n examples and corresponding y_i labels. Prototypes ω^j with labels $c(\omega^j)$ are iteratively updated as below;

$$\omega^j := \begin{cases} \omega^j + \eta \cdot (\xi_i - \omega^j) & \text{if} \quad y_i = c(\omega^j) \\ \omega^j - \eta \cdot (\xi_i - \omega^j) & \text{otherwise} \end{cases} \tag{1}$$

$$\xi \leftarrow c(\omega^j) \quad \text{where} \quad \arg\min_{\xi} d(\xi, \omega^j) \tag{2}$$

- Sato and Yamada[Sato and Yamada, 1995] introduce a cost function for boundary optimization in **Generalised LVQ**(GLVQ).
  The cost function is defined as;

$$E = \sum_{i}^{n} \Phi(\mu_i) \quad \text{where} \quad \mu_i = \frac{d(\xi_i, \omega^J) - d(\xi_i, \omega^K)}{d(\xi_i, \omega^J) + d(\xi_i, \omega^K)} \tag{3}$$

-

$\mu_i \in [-1, 1]$ shows the confidence of the classification.

Relevance learning

Optimize similarity measure by adapting to target dataset.

$$d^{\lambda}(\xi, \omega) = \sum_{i=1}^{N} \lambda_i (\xi_i - \omega_i)^2 \tag{4}$$

Weight factors λ_i as known as ***relevances*** are iteratively updated for each learning step[Bojer et al., 2001].
Is a substitute to general Euclidean distance and increases classification performance.

`Generalized Relevance LVQ`(GRLVQ)[Hammer and Villmann, 2002] combines both concepts of relevance learning and a cost function as in equations 3 and 4.

- *Global Metric Tensors*

A global full transformation matrix is defined to be adapted by stochastic gradient descent for each learning step.

Generalized Matrix LVQ(GMLVQ)[Schneider et al., 2009b, Schneider et al., 2009a] implements this metric.

$$d^{\Lambda}(\xi, \omega) = (\xi - \omega)^T \Lambda (\xi - \omega) \tag{5}$$

where Λ is full $N \times N$ matrix that represents the correlations between the feature dimensions. It is derivable to Euclidean distance in a transformed space as shown below;

$$\begin{aligned} \Lambda &= \Omega^T \Omega \\ d^{\Lambda}(\xi, \omega) &= (\xi - \omega)^T \Omega^T \Omega (\xi - \omega) \\ d^{\Lambda}(\xi, \omega) &= [\Omega^T (\xi - \omega)]^2 \end{aligned} \tag{6}$$

The diagonal of Λ corresponds to the relevance values of the feature dimensions.

- *Local Metric Tensors*

Learn the localized dissimilarities w.r.t to the distinct class prototypes.
Define non-linear decision boundaries with the local transformation matrix.

Localized GMLVQ(LGMLVQ)[Schneider et al., 2009b]- direct extension of GMLVQ.

$$d^{\Lambda^j}(\xi, \omega^j) = (\xi - \omega^j)^T \Lambda^j (\xi - \omega^j) \tag{7}$$

where $\Lambda^j = \Omega^{jT}\Omega^j$ and j is a given prototype.

$$E_{LGMLVQ} = \sum_i^n \Phi(\mu^i_{local}) \quad \text{where} \quad \mu^i_{local} = \frac{d_J^{\Lambda^j} - d_K^{\Lambda^k}}{d_J^{\Lambda^j} + d_K^{\Lambda^k}} \tag{8}$$

where $d_J^{\Lambda^j}$ and $d_K^{\Lambda^k}$ are the distances of the points ξ^i from the closest correct and incorrect prototypes respectively.

- *Local Metric Tensors*

Localized Limited Rank Matrix
LVQ(LLiRAM)[Bunte et al., 2012, Bunte, 2011]
Transform data to a low dimensional representation within the algorithm.
The distance metric is defined as;

$$d_j^{\Psi^j}(\xi, \omega^j) = (\xi - \omega^j)^T \Omega^T \Psi^{jT} \Psi^j \Omega (\xi - \omega^j) \tag{9}$$

where Ω is a rectangular $M \times N$ matrix of the global linear dimensionality reduction and Ψ^j is the local dissimilarity in the transformed space for each prototype ω^j.

Experiments and Discussion

Experimental Setup

- **`Baseline Setup`** - Without resampling techniques
  10-fold cross validation with a ***`90/10`*** class-wise split, Z-score transformation, 1 prototype per class, regularization set to 0.00001.
- **`Resampled setup`**
  Involves resampling techniques of SMOTE or Borderline-SMOTE after Z-score transformation with the 10-cross validation folds.

Experimental Setup BV1 - *data1*

	GMLVQ	LGMLVQ2	LGMLVQ3
Precision 2	0 ± 0	0.586 ± 0.505	0.754 ± 0.402
Specificity 2	1 ± 0	0.9997 ± 0.0007	0.999 ± 0.0012
TPR 1	0.989 ± 0.0058	0.99 ± 0.0071	0.99 ± 0.0075
TPR 2	0 ± 0	0.431 ± 0.3903	0.544 ± 0.311
TPR 3	0.997 ± 0.0028	0.997 ± 0.0030	0.997 ± 0.0028
Avg training error	0.031 ± 0.00037	0.0202 ± 0.0093	0.016 ± 0.0079
Avg test error	0.0312 ± 0.00262	0.0198 ± 0.0113	0.017 ± 0.0067

Table 1: *BV1* results showing precision and specificity of the minority class, TPR 1,2 and 3 represent the True Positive Rate/recall of the corresponding classes 1, 2 and 3, average training and test error all with corresponding standard deviation. These results are for the methods GMLVQ, LGMLVQ2 i.e. with rank 2 and LGMLVQ3 i.e with rank 3 for experimental setup **BV1**.

Experimental Setup RV1

	GMLVQ	LGMLVQ2	LGMLVQ3	LLiRaM
Precision 2	0.451 ± 0.108	0.698 ± 0.0213	0.706 ± 0.092	0.52 ± 0.071
Specificity 2	0.968 ± 0.013	0.989 ± 0.0011	0.989 ± 0.0048	0.9775 ± 0.007
TPR 1	0.949 ± 0.019	0.987 ± 0.001	0.986 ± 0.0069	0.9595 ± 0.016
TPR 2	0.913 ± 0.060	0.934 ± 0.0079	0.938 ± 0.059	0.9 ± 0.0672
TPR 3	0.973 ± 0.012	0.985 ± 0.0015	0.984 ± 0.0063	0.981 ± 0.006
Avg training error	0.037 ± 0.005	0.016 ± 0.0013	0.0155 ± 0.001	0.0422 ± 0.002
Avg test error	0.037 ± 0.12	0.017 ± 0.0065	0.0167 ± 0.006	0.0282 ± 0.007

Table 2: *RV1* results for the methods GMLVQ, LGMLVQ2 with rank 2 , LGMLVQ3 with rank 3 and LLiRaM for experimental setup **RV1**. The data in these experiments is all resampled based on SMOTE.

Experimental Setup BV2 - *data2*

	LGMLVQ2	LGMLVQ3
Precision 2	0.9544 ± 0.0234	0.9611 ± 0.0283
Specificity 2	0.9972 ± 0.0015	0.9976 ± 0.0017
TPR 1	0.9826 ± 0.0059	0.9833 ± 0.0058
TPR 2	0.8176 ± 0.0515	0.8157 ± 0.0508
TPR 3	0.9902 ± 0.0042	0.9904 ± 0.0041
Avg training error	0.0232 ± 0.001	0.023 ± 0.00058
Avg test error	0.02396 ± 0.005	0.0237 ± 0.0049

Table 3: BV2 results showing precision and specificity of the minority class, TPR 1,2 and 3 represent the True Positive Rate/recall of the corresponding classes 1, 2 and 3, average training and test error all with corresponding standard deviation. These results are for the methods *LGMLVQ2* with rank 2 and *LGMLVQ3* i.e. with rank 3 for experimental setup **BV2**.

Experimental Setup RV2

	LGMLVQ 2	LGMLVQ3	LGMLVQ2 -BL
Precision 2	0.8362 $\pm$ 0.0461	0.8449 $\pm$ 0.0456	0.697 $\pm$ 0.0631
Specificity 2	0.9868 $\pm$ 0.0043	0.9877 $\pm$ 0.0043	0.97 $\pm$ 0.0097
TPR 1	0.967 $\pm$ 0.0098	0.9674 $\pm$ 0.0093	0.922 $\pm$ 0.0264
TPR 2	0.939 $\pm$ 0.0457	0.9392 $\pm$ 0.0457	0.949 $\pm$ 0.0482
TPR 3	0.9818 $\pm$ 0.0043	0.9825 $\pm$ 0.0043	0.98 $\pm$ 0.0056
Avg training error	0.0261 $\pm$ 0.001	0.02593 $\pm$ 0.001	0.043 $\pm$ 0.007
Avg test error	0.0264 $\pm$ 0.0067	0.0259 $\pm$ 0.0067	0.0435 $\pm$ 0.008

Table 4: *RV2* results are for the methods LGMLVQ2 with rank 2 , LGMLVQ3 with rank 3 and LGMLVQ2 -BL for experimental setup **RV1**. LGMLVQ2-BL is resampled using Borderline-SMOTE technique and the rest by SMOTE

Experimental Setup RV2- -*LGMLVQ2*

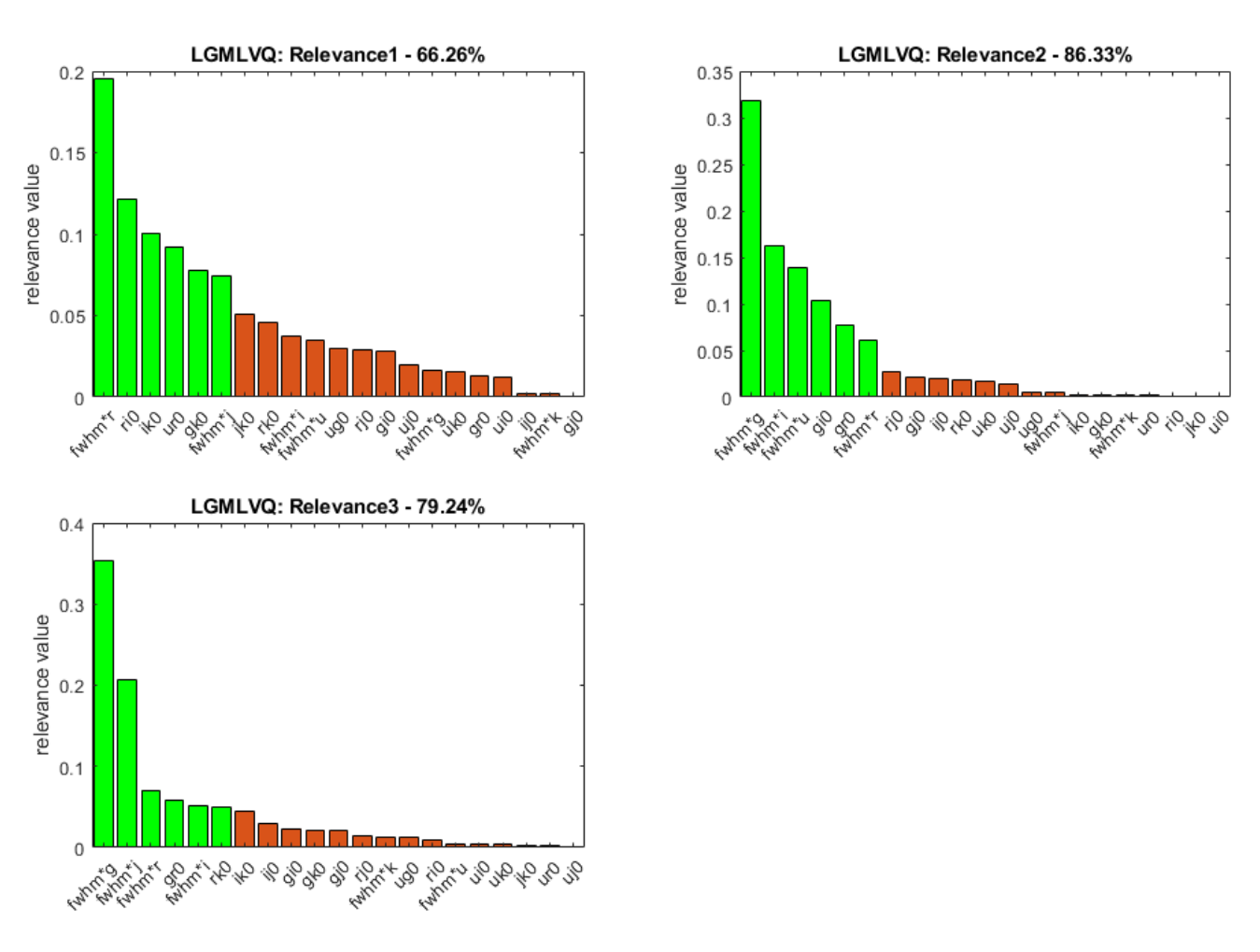


Figure 4: Top 6 class-wise relevant features in the normalized profile for LGMLVQ2 with the features' corresponding percentage of contribution

Experimental Setup RV2 -*LGMLVQ2*

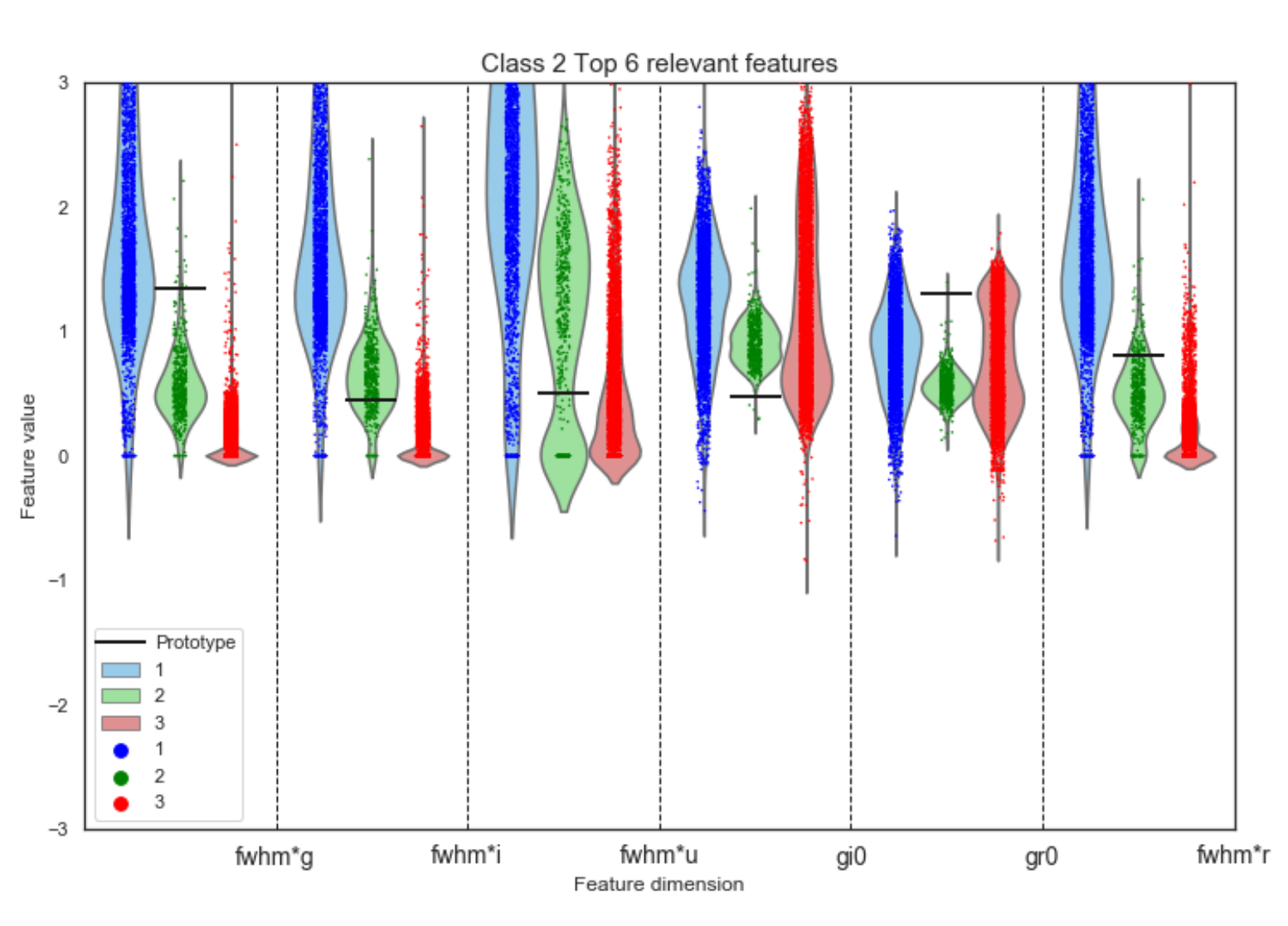
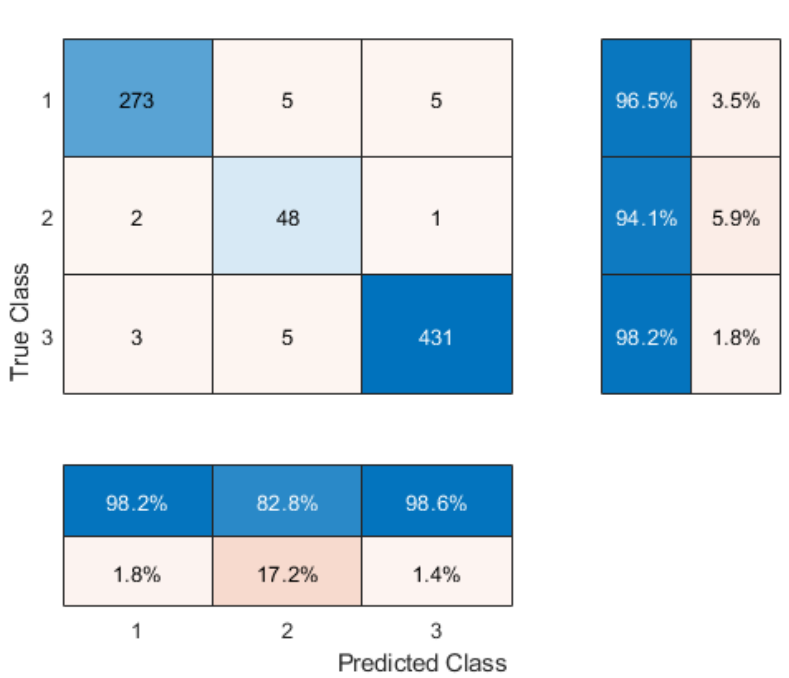
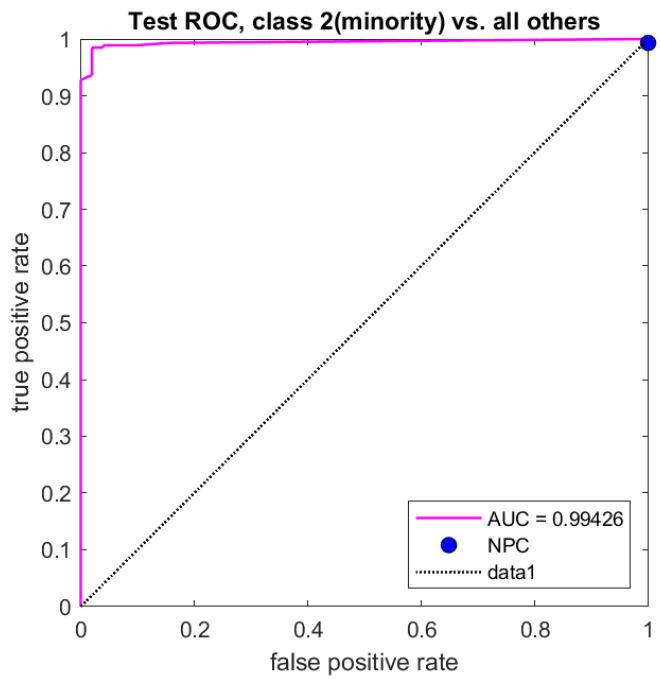


Figure 5: Further illustration of the Top 6 relevant features' prototype positions relative to the feature value distribution

Experimental Setup RV2 -*LGMLVQ2*



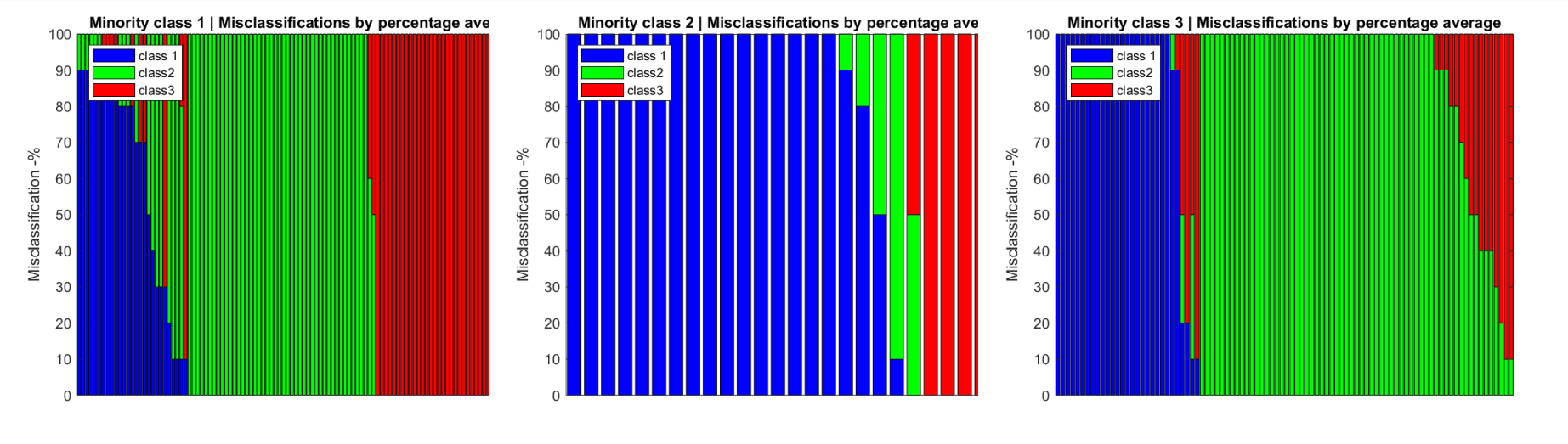
(a)



(b)

Figure 6: (a) is the test confusion matrix for the model *LGMLVQ2* in *RV2* (b) shows the corresponding ROC curve of minority class 2 vs all the other classes and the corresponding AUROC value of 0.99387

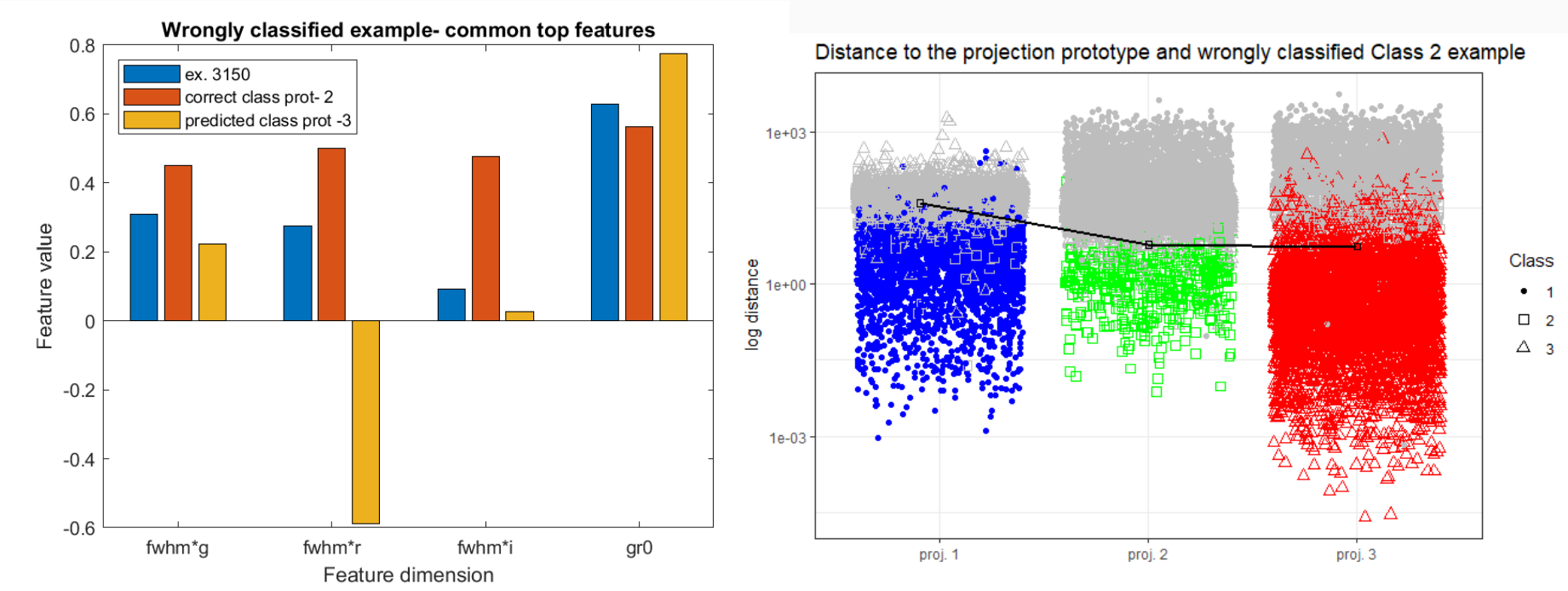
Experimental Setup RV2 -*LGMLVQ2*



(a) (b) (c)

Figure 7: Misclassification rates for the wrongly classified examples in each class representing which classes there are misclassified for. (a) 64 of the 125 misclassified class 1 examples are often identified as class 2. (b) 20 of the 27 class 2 examples are often classified as class 1. (c) 66 of the 92 class 3 examples are classified as class 2

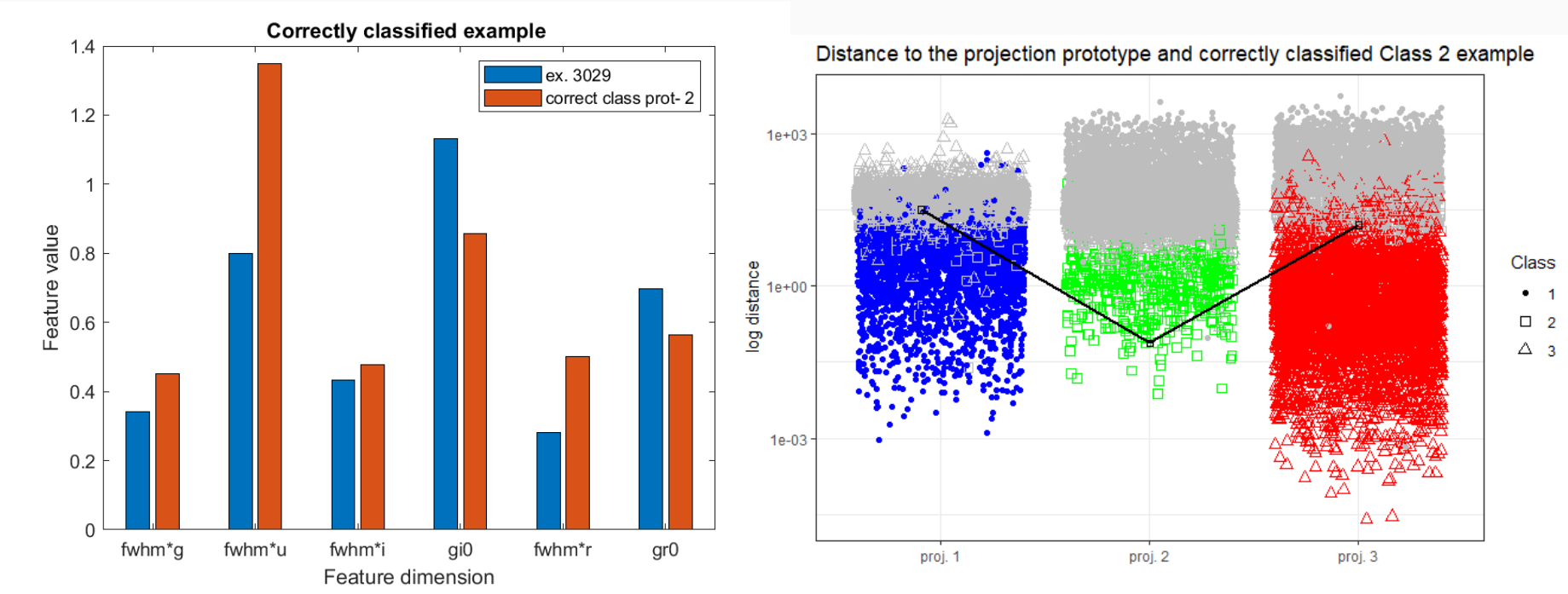
Wrongly classified example -*LGMLVQ2*



(a) (b)

Figure 8: 100% Wrongly classified example from the minority class 2. (a) Common top feature positions of the examples relative to the closest correct and wrong prototypes (b) distances to the projection prototype and arc representation of the same example relative to the other observations in each class.

Correctly classified example -*LGMLVQ2*



(a) (b)

Figure 9: 100% Correctly classified example from the minority class 2. (a) Top 6 feature positions of the examples relative to the closest correct prototype (b) arc representation for same example in distance to projection prototype.

Summary of discussion

- Angular size features - $fwhm^*g, fwhm^*i, fwhm^*u$ are predominantly important for the minority class than expected.
- Sizes are expected to be used to discriminate class 1 from class 2 and 3.
- Class 2 objects are faint hence measurements can be larger than actual size. This introduces a bias in the data.
- Astronomically more relevant features of *ui*- age of object and *ik* - metallicity do not vie well for the minority class. Observations in the *u* and *K* filters are hard and have a lower signal to noise ratio. Also other color indices like gi, gr, rk etc. partially carry the information.
- Appropriate pre-processing is crucial for this research
- The class boundaries are evidently non-linear and local metric tensor method of LGMLVQ is appropriate for this data

Future Work

Possible Future work

- Use trained models on novel observation to aid further research into the Ultra-compact Dwarf galaxies(UCDs) and Globular Clusters.
- More work on preliminary feature selection to address the disparity from astronomical expectations.
- Explore missing value imputation techniques to cater for information loss
- Angle LVQ[Bunte et al., 2016] that defines angular based similarity and minimises within-class variation.

Questions?

References i


Bertin, E. and Arnouts, S. (1996).
Sextractor: Software for source extraction.
Astronomy and Astrophysics Supplement Series, 117(2):393–404.


Bojer, T., Hammer, B., Schunk, D., and von Toschanowitz, K. T. (2001).
Relevance determination in learning vector quantization.
In *PROC. OF EUROPEAN SYMPOSIUM ON ARTIFICIAL NEURAL NETWORKS*.


Bunte, K. (2011).
Adaptive dissimilarity measures, dimension reduction and visualization.
PhD thesis.

References ii


Bunte, K., Baranowski, E. S., Arlt, W., and Tino, P. (2016).
Relevance learning vector quantization in variable dimensional spaces.
In Hammer, B., Martinetz, T., and Villmann, T., editors, *New Challenges in Neural Computation (NC^2)*, Workshop of the GI-Fachgruppe Neuronale Netze and the German Neural Networks Society in connection to GCPR 2016, pages 20–23, Hannover, Germany.


Bunte, K., Schneider, P., Hammer, B., Schleif, F.-M., Villmann, T., and Biehl, M. (2012).
Limited rank matrix learning, discriminative dimension reduction and visualization.
Neural Networks, 26:159 – 173.


Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002).
Smote: Synthetic minority over-sampling technique.
J. Artif. Int. Res., 16(1):321–357.

References iii


Hammer, B. and Villmann, T. (2002).
Generalized relevance learning vector quantization.
Neural Networks, 15(8):1059 – 1068.


Han, H., Wang, W.-Y., and Mao, B.-H. (2005).
Borderline-smote: A new over-sampling method in imbalanced data sets learning.
In Huang, D.-S., Zhang, X.-P., and Huang, G.-B., editors, *Advances in Intelligent Computing*, pages 878–887, Berlin, Heidelberg. Springer Berlin Heidelberg.


Kohonen, T. (1997).
***Learning Vector Quantization*, pages 175–189.**
Springer Berlin Heidelberg, Berlin, Heidelberg.


Munoz, R. P., Puzia, T. H., Lançon, A., Peng, E. W., Cote, P., Ferrarese, L., Blakeslee, J. P., Mei, S., Cuillandre, J.-C., Hudelot, P., et al. (2013).
The next generation virgo cluster survey-infrared (ngvs-ir). i. a new near-ultraviolet, optical, and near-infrared globular cluster selection tool.
The Astrophysical Journal Supplement Series, 210(1):4.


Sato, A. and Yamada, K. (1995).
Generalized learning vector quantization.
In *Proceedings of the 8th International Conference on Neural Information Processing Systems*, NIPS'95, page 423–429, Cambridge, MA, USA. MIT Press.


Schneider, P., Biehl, M., and Hammer, B. (2009a).
Adaptive relevance matrices in learning vector quantization.
Neural Comput., 21(12):3532–3561.

Schneider, P., Biehl, M., and Hammer, B. (2009b).
Distance learning in discriminative vector quantization.
Neural Computation, 21(10):2942–2969.

Stetson, P. B. (2013).
***Astronomical Photometry*, pages 1–34.**
Springer Netherlands, Dordrecht.