ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Prototype-based Analysis of GAMA Galaxy Catalogue Data

Aleke Nolte[1]*, Lingyu Wang[2], and Michael Biehl[1]

1 - Univ. of Groningen - Johann Bernoulli Inst. for Mathematics and Computer Science
P.O. Box 407, 9700 AK Groningen - The Netherlands

2- SRON Netherlands Institute for Space Research
and University of Groningen - Katpeyn Astronomical Institute
Landleven 12 - 9747 AD Groningen - The Netherlands

Abstract. We present a prototype-based machine learning analysis of labeled galaxy catalogue data containing parameters from the Galaxy and Mass Assembly (GAMA) survey. Using both an unsupervised and supervised method, the Self-Organizing Map and Generalized Relevance Matrix Learning Vector Quantization, we find that the data does not fully support the popular visual-inspection-based galaxy classification scheme employed to categorize the galaxies. In particular, only one class, the *Little Blue Spheroids*, is consistently separable from the other classes. In a proof-of-concept experiment, we present the galaxy parameters that are most discriminative for this class.

1 Introduction

Telescope images of galaxies reveal a multitude of appearances, ranging from disk-like galaxies over galaxies with spiral arms to more irregular shapes. The study of galaxy morphological classification plays an important role in astronomy: The spatial distribution of galaxy types provides valuable information for the understanding of galaxy formation and evolution.
The assignment of morphological classes to observed galaxies is a task which is commonly handled by astronomers. As manual labeling of galaxies is time consuming and expert-devised classification schemes may be subject to cognitive biases, machine learning techniques have great potential to advance astronomy by: 1) investigating automatic classification strategies, and 2) by evaluating to which extent existing classification schemes are supported by the observed data.
In this work, we want to make a contribution along both lines by analyzing a galaxy catalogue which has been annotated using a popular classification scheme proposed by Kelvin [1]. To this end, we apply both an unsupervised and a supervised prototype-based method. In our analysis, we first assess whether Kelvin's scheme is consistent with a clustering of the data generated by the unsupervised Self Organizing Map (SOM) [2]. We then investigate if the morphological classification can be reproduced by Generalized Relevance Matrix Learning Vector Quantization (GMLVQ) [3], a powerful supervised prototype-based method. In addition to providing an evaluation of the classification scheme via the proxy of classification performance (high accuracies would indicate a good separation of classes), GMVLQ also allows to identify

*We acknowledge financial support by the EU's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 721463 to the SUNDIAL ITN network.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

the feature dimensions that are of particular relevance for a classification problem. As we find the employed classification scheme to be not fully supported by the data, we present the parameters that are relevant for distinguishing the most clearly pronounced class, the *Little Blue Spheroids (LBS)*, from the remaining classes.

2 Data

In this work we analyze a labeled galaxy catalogue [4] containing 42 parameters which have been derived from spectroscopic and photometric observations (measurements of flux intensities in different wavelength bands) from the Galaxy and Mass Assembly (GAMA) survey [5] for 7941 astronomical objects. For each galaxy, a class label has been determined by astronomers following a visual inspection based classification scheme described by Kelvin et al. [1]. The scheme assigns galaxies to 9 classes: 1-*Ellipticals* (11%), 2-*Little blue spheriods* (11%), 3-*Early-type spirals* (10%), 4-*Early-type barred spirals* (1%), 5-*Intermediate-type spirals* (15%), 6-*Intermediate-type barred spirals* (2%), 7-*Late-type spirals & Irregulars*(45%), 8-*Artefacts* (0.4%) and 9-*Stars* (0.005%). Numbers in parantheses represent the class-wise prevalences in the available data. We will refer to the classes by their class index (1-9). We exclude the *LOG-SurfaceDensityErr* parameter due to numerous missing measurements, and remove samples with missing measurements in any of the remaining features, resulting in a final dataset of 7356 astronomical objects of dimensionality $n=41$.

3 SOM Analysis

The self-organizing map (SOM) [2, 6] is an unsupervised prototype-based method which allows to generate topology-preserving low-dimensional representations of high-dimensional data. In the default formulation, a SOM is composed of map units that are arranged in a two-dim. lattice. Each unit has a defined set of neighbors which are influenced when the unit undergoes changes. A *prototype* of input dimensionality is associated with each unit of the map and the training process follows a *competitive learning* paradigm: For each data point $\boldsymbol{\xi}$, the best matching unit (BMU) $\mathbf{m}_{c(\xi)}$ is determined, and subsequently only $\mathbf{m}_c$ and its neighbors are modified by the training procedure. The BMU $\mathbf{m}_{c(\xi)}$ is given by the map unit with the prototype closest to $\boldsymbol{\xi}$ and therefore fulfills

$$\forall i, \|\boldsymbol{\xi}-\mathbf{m}_c\| \leq \|\boldsymbol{\xi}-\mathbf{m}_i\|.$$

In the SOM batch learning rule ([7]), the map units $\mathbf{m}_i$ are updated following

$$\mathbf{m}_i(t+1)=(\textstyle\sum_{j=1}^{m} h_{j,i}\,\mathbf{s}_j)/(\sum_{j=1}^{m} n_{V_j} h_{j,i}) \text{,with } \mathbf{s}_i=\textstyle\sum_{\boldsymbol{\xi}_k \in V_i} \boldsymbol{\xi}_k\,,$$

where m is the number of map units, V_i is the set of data points with BMU $\mathbf{m}_i(t)$, n_{V_i} is the cardinality of this set and $h_{i,j}$ is the neighborhood function determining the strength of the influence of unit i to j.
We perform an analysis of the galaxy catalogue data using the SOM implementation "SOM toolbox for Matlab5" [7]. Due to the SOM's sensitivity to outliers we normalize the data using a centered logistic normalization:

$$\xi'=1/(1+e^{-\hat{\xi}})-0.5 \text{ with } \hat{\xi}=(\xi-\bar{\xi})/\sigma_\xi\,,$$

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

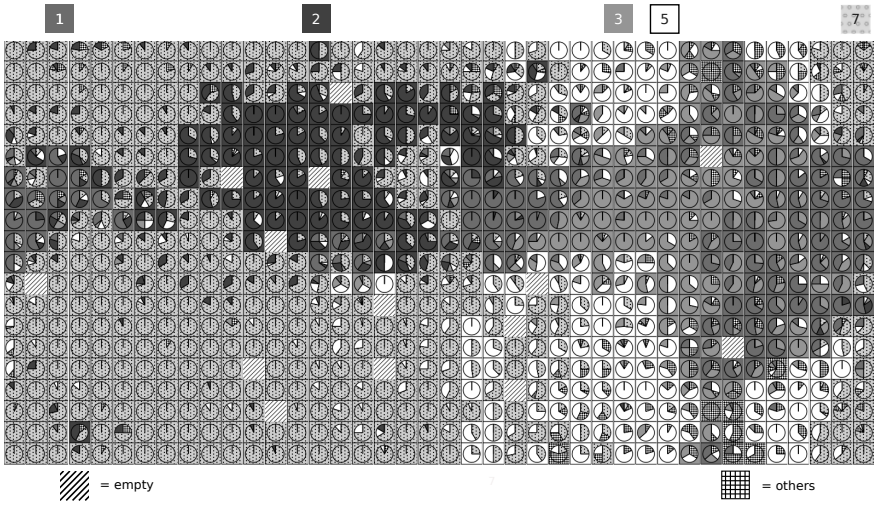

Fig. 1: SOM visualization displaying an unsupervised clustering of galaxy data from GAMA. The SOM is post-labeled using galaxy-class information, where the pie chart for each unit encodes a histogram of class-wise hits. A unit's background color indicates the class with the majority of hits. A color version of this plot can be found at `http://doi.org/10.5281/zenodo.1092761`.

where $\bar{\xi}$ and σ_ξ are the empirical mean and standard deviation of the feature.
We initialize the SOM by spacing the prototypes regularly in the coordinate system given by the two largest eigenvectors of the data matrix [7]. Based on preliminary experiments we choose a SOM of size 20×40 units with a rectangular neighborhood function and toroidal topology and train using the option *long* of the default toolbox settings. After training, the galaxy label information is used to display the specificity of a map unit via unit-wise histograms of *hits* (pie charts in Figure 1), i.e. the distribution of the class labels of the data points that have a particular map unit as BMU. Additionally, the class with the majority of hits is indicated by the background color of each pie plot.
In the pie charts of Fig. 1 it can be seen that there is no pronounced separation of classes and that many units respond to samples from more than one class. Classes 1 and 3 are particularly intermixed. However, class 2 (LBS) forms a fairly compact cluster. We note that the basic structure remains similar for larger or smaller maps, or when the over-represented class 7 is under-sampled.

4 GMLVQ analysis

Generalized Relevance Matrix LVQ (GMLVQ) [3] is an extension of Learning Vector Quantization (LVQ) [6]. LVQ is a supervised prototype-based method, in which prototypes are annotated with a class label. The prototypes are adapted based on

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

the label information of the training data: If the BMU is of the same class as a given data point, the prototype is moved towards the data point, while in the case of a BMU with incorrect class label, the prototype is repelled. While both SOM and LVQ assess similarities between prototypes and data points using the Euclidean distance, GMLVQ learns a distance measure that is tailored to the data, allowing it to suppress noisy feature dimensions or to emphasize distinctive features and their pair-wise combinations. GMLVQ therefore considers a generalized distance

$$d^{\Lambda}(\mathbf{w},\boldsymbol{\xi}) = (\boldsymbol{\xi}-\mathbf{w})^T \Lambda (\boldsymbol{\xi}-\mathbf{w}) \quad \text{with } \Lambda = \Omega^T \Omega \quad \text{and} \quad \textstyle\sum_i \Lambda_{ii} = 1,$$

where Λ is an $n \times n$ a positive semi-definite distance matrix, and $\mathbf{w}$ is a prototype. The parameters $\{\mathbf{w}_i\}$ and Λ are optimized based on a heuristic cost function, see [3].

$$E_{\mathrm{GMLVQ}} = \textstyle\sum_{i=1}^{P} \mu_i^{\Lambda}, \text{ with } \mu_i^{\Lambda} = (d_J^{\Lambda}(\boldsymbol{\xi}_i) - d_K^{\Lambda}(\boldsymbol{\xi}_i))/(d_J^{\Lambda}(\boldsymbol{\xi}_i) + d_K^{\Lambda}(\boldsymbol{\xi}_i)),$$

where $d_J^{\Lambda}(\boldsymbol{\xi}) = d_J^{\Lambda}(\mathbf{w}_J, \boldsymbol{\xi})$ denotes the distance to the closest correctly labeled prototype, and $d_K^{\Lambda}(\boldsymbol{\xi}) = d_K^{\Lambda}(\mathbf{w}_K, \boldsymbol{\xi})$ denotes the distance to the closest incorrect prototype. If the closest prototype has an incorrect label, $d_K^{\Lambda}(\boldsymbol{\xi}_i)$ will be smaller than $d_J^{\Lambda}(\boldsymbol{\xi}_i)$, hence, the corresponding μ_i^{Λ} is positive. Minimization of E_{GMLVQ} will therefore favor the correctness of nearest prototype classification. In a stochastic gradient descent procedure based on single examples the update reads

$$\mathbf{w}_{J,K} \leftarrow \mathbf{w}_{J,K} - \eta_w \partial \mu_i / \partial \mathbf{w}_{J,K} \text{ and } \Omega \leftarrow \Omega - \eta_\Omega \partial \mu_i / \partial \Omega . \tag{1}$$

Derivations and full update rules can be found in [3].

To assess relevances of features and discriminability between classes we train and evaluate GMLVQ on the galaxy catalogue data making use of a publicly available implementation which employs batch gradient descent [8]. To maintain consistency, we normalize the data using logistic normalization. As the GMLVQ cost function is implicitly biased toward classes with larger numbers of samples, we disregard the classes which contain only few samples (classes 4,6,8,9) and train and evaluate the classifier on size-balanced random subsets from the remaining five larger classes. For our experiments, we specify one prototype per class and run the algorithm for 100 epochs using the implementation's default settings. We validate the algorithm by performing a class-balanced *repeated random sub-sampling validation* by randomly selecting 743 datapoints (the cardinality of the smallest class, class 3) per class for each validation run, for a total of 50 runs. For each sub-sampled data set, both the training (3343 samples, 90%) and validation set (371 samples (10%)) are class-balanced. The resulting relevances and confusion matrix (both averaged over all validation runs) are displayed in Figure 2 and Table 4. The confusion matrix corroborates the findings from the SOM analysis: The classifier has difficulty distinguishing classes 1 and 3, placing about 19% of samples from class 1 into class 3 and vice versa. Also classes 5 and 7 can only be classified with limited accuracies of 73% and 71%. In line with the compactness of its representation by the SOM, only class 2 (*little blue spheroids*, LBS), are classified with a high accuracy (90.7%). Both the lack of pronounced clusters in the SOM and the moderate classification performance of GMLVQ may be an indication for the presence of label noise that is

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

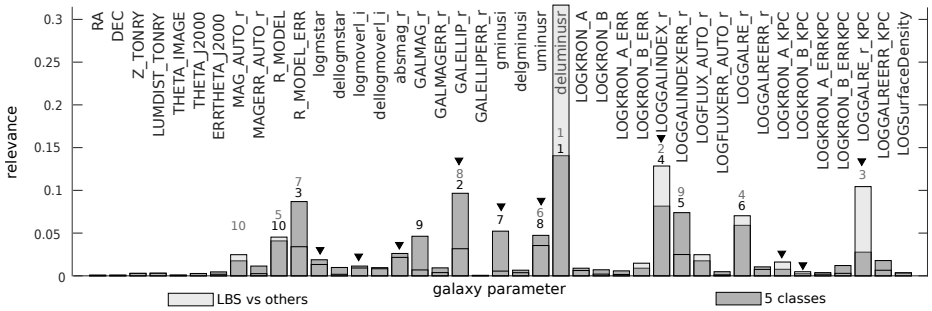

Fig. 2: Average GMVLQ feature relevances for the task of distinguishing the 5 largest galaxy classes (dark) and the little blue spheroid class from the other 4 classes (light). The 10 most relevant features are indicated by digits (dark/light) corresponding to the rank of the feature. Arrows indicate features selected in [4]. Explanations of the galaxy parameters and additional references can be found at `http://doi.org/10.5281/zenodo.1092761`.

	1	2	3	5	7
1	63.8	10.8	19.3	5.5	0.7
2	3.1	90.7	0.1	1.3	4.8
3	18.4	1.7	66.1	13.5	0.3
5	2.0	6.8	9.3	73.7	8.2
7	1.1	13.1	0.1	14.6	71.1

Table 1: Averaged GMVLQ confusion matrix displaying the percentage of predicted class labels (columns) for samples from each class (rows) for the 5 largest galaxy classes.

rooted in the subjective nature of galaxy classification [4, 9]. It may also indicate that the classification scheme is not fully supported by the data. As the relevance profile reflects the discriminability of parameters only within this possibly ill-defined setting, we additionally consider the relevance profile for the less ambiguous sub-problem of separating LBS from other galaxies. Following the procedure described above (with 1512 training and 168 validation samples per validation run), we find that LBS can be well distinguished (AUC(ROC)=0.96) from the other classes. Furthermore, in the two-class setting 8 of the 10 most relevant features are also among the 10 most relevant ones found for the full five-class setting. The coinciding parameters are *GALELLIP_r*, *LOGGALINDEXERR_r*, *LOGGALINDEX_r*, *LOGGALRE_r*, *R_MODEL*, *R_MODEL_ERR*, *deluminusr*, and *uminusr*.

5 Discussion

The results presented above suggest that there may be some inconsistencies in the investigated morphological classification scheme: The clustering produced by the unsupervised SOM is only moderately consistent with the galaxy classes and it has proven difficult to distinguish galaxy types using supervised GMLVQ. In our analysis, class 1 (*Ellipticals*) and 3 (*Early-type spirals*) are particularly difficult to differentiate, while class 2 (LBS) seems to be well separable.
The difficulty of training a successful classifier was also observed in [4], where class-wise averaged accuracies are also around 75%. Possible explanations for poor classification

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

performance may be lack of discriminative power of the employed classifiers, mis-labelings of certain galaxies (a possibility already discussed in [4]), or that essential parameters are not contained in the data set. In the first case, employing even more flexible classifiers, e.g. GMLVQ with local relevance matrices [3], may improve the classification performance. In the second case, if mis-labelings are restricted to "neighboring" classes in an assumed underlying class ordering, ordinal classification may provide further insights [10]. Yet, our results do not rule out the possibility that the true, underlying grouping of galaxies is considerably different and less clear-cut than the investigated one. Further data-driven analyses of galaxy parameters and images with advanced clustering methods might reveal alternative groupings.
In [4], 10 of the 42 parameters were selected manually. Out of GMLVQ's 8 most relevant features, 4 overlap or highly correlate with the 10 features selected in [4], namely *GALELLIP_r, LOGGALINDEX_r, uminus_r, and LOGGALRE_r*. These parameters are related to the ellipticity, light distribution, size and color of the galaxy. Note that the non-overlapping features comprise model fitting errors (*LOGGALINDEXERR_r, R_MODEL_ERR, deluminusr*), indicating that a galaxy's model consistency varies over morphological classes.

Acknowledgements GAMA is a joint European-Australasian project based around a spectroscopic campaign using the Anglo- Australian Telescope. The GAMA input catalogue is based on data taken from the Sloan Digital Sky Survey and the UKIRT Infrared Deep Sky Survey. Complementary imaging of the GAMA regions is being obtained by a number of in- dependent survey programmes including GALEX MIS, VST KiDS, VISTA VIKING, WISE, Herschel-ATLAS, GMRT and ASKAP providing UV to radio coverage. GAMA is funded by the STFC (UK), the ARC (Australia), the AAO, and the participating institutions. The GAMA website is `http://www.gama-survey.org/`. We thank Sreevarsha Sreejith and Lee Kelvin for helpful feedback and discussions.

References

[1] L.S. Kelvin, S.P. Driver, A. SG Robotham, et al. Galaxy and mass assembly (gama): ugrizyjhk sérsic luminosity functions and the cosmic spectral energy distribution by hubble type. *Monthly Notices of the Royal Astronomical Society*, 439(2):1245–1269, 2014.

[2] T. Kohonen. The self-organizing map. *Neurocomputing*, 21(1):1–6, 1998.

[3] P. Schneider, M. Biehl, and B. Hammer. Adaptive relevance matrices in Learning Vector Quantization. *Neural Computation*, 21(12):3532–3561, 2009.

[4] S. Sreejith, S. Pereverzyev Jr., L. S. Kelvin, et al. Galaxy and mass assembly: Automatic morphological classification of galaxies using statistical learning. in preparation.

[5] J. Liske, IK Baldry, S.P. Driver, et al. Galaxy and mass assembly (gama): end of survey report and data release 2. *Monthly Notices of the Royal Astronomical Society*, 452(2):2087–2126, 2015.

[6] T. Kohonen. *Self-organizing maps*. Springer, 1997.

[7] J. Vesanto, J. Himberg, E. Alhoniemi, and Juha Parhankangas. Som toolbox for matlab 5. *Helsinki University of Technology, Finland*, 2000.

[8] M. Biehl. Gmlvq demo code. http://www.cs.rug.nl/~biehl/gmlvq. Last accessed: 2017-11-21.

[9] A Dressier, JP Huchra, S van den Bergh, and S Raychaudhury. Galaxies, human eyes, and artificial neural networks. *Science*, 267:859, 1995.

[10] S. Fouad and P. Tino. Adaptive metric learning vector quantization for ordinal classification. *Neural Computation*, 24(11):2825–2851, 2012.