Astronomical Data Analysis Software and Systems: XXIV
ASP Conference Series, Vol. 495
A. R. Taylor and E. Rosolowsky, eds.

Automatic Galaxy Classification via Machine Learning Techniques: Parallelized Rotation/Flipping INvariant Kohonen Maps (PINK)

Kai Lars Polsterer[1], Fabian Gieseke[2], and Christian Igel[2]

[1]*HITS gGmbH, Astroinformatics, Heidelberg, Germany*

[2]*University of Copenhagen, Department of Computer Science, Denmark*

Abstract. In the last decades more and more all-sky surveys created an enormous amount of data which is publicly available on the Internet. Crowd-sourcing projects such as Galaxy-Zoo and Radio-Galaxy-Zoo used encouraged users from all over the world to manually conduct various classification tasks. The combination of the pattern-recognition capabilities of thousands of volunteers enabled scientists to finish the data analysis within acceptable time. For up-coming surveys with billions of sources, however, this approach is not feasible anymore. In this work, we present an unsupervised method that can automatically process large amounts of galaxy data and which generates a set of prototypes. This resulting model can be used to both visualize the given galaxy data as well as to classify so far unseen images.

1. Motivation

Today, the manual inspection by an expert of all objects in the available large-scale astronomical databases is impossible. Caused by the exponential growth in size and complexity of data-sets in astronomy, new explorative analysis methods are required. For a few of the current surveys, volunteers have addressed some tasks quite successfully (e.g., classification). The Galaxy Zoo project is a good example of how to make use of more than 100,000 volunteers to derive a morphological analysis for about 900,000 galaxies (Lintott et al. 2011). However, the projected increase in the number of objects for the next generation of all-sky survey missions renders such a manual inspection impossible. Another challenge is that volunteers might not have the required background to detect rare and interesting objects. Each crowd-source project is laid out to deal with some specific task. Thus, new scientific questions/data-sets could require the definition of new crowd-source projects. Therefore new methods need to be developed that combine semi-automatic data analysis schemes with the visual recognition capabilities, the creativity, and keen perception of the human brain. By using computers to pre-process and pre-analyze the data, we try to assist astronomers to conduct such tasks in a semi-automatic manner instead of a fully manual analysis via crowd-sourcing projects. Similar and frequent objects can be combined/sorted by machine learning models, which yield only a single representative that needs manual inspection by the scientist. The goal of our work is to enable astronomers to efficiently perform a morphological analysis on huge amounts of pre-processed data (e.g., images or radio-synthesis data). Concerning the morphological taxonomy of galaxies, Hubble's tuning-fork diagram provides a popular approach to organize the observed classes. Within its limitations, this diagram is a nice way of presenting the different classes of galaxies. Besides, this typical diagram depicts a topological sorting of the different galaxy classes.

In the past, dimensionality reduction techniques that are able to compute topological maps, i.e., latent embeddings, haven shown good results (Kramer et al. 2013) based on images of galaxies from the Sloan Digital Sky Survey (SDSS) (Ahn et al. 2014). Those dimension reduction techniques aim at projecting complex, high-dimensional data to a low-dimensional feature spaces while preserving similarities and neighborhood relations between the original data points. This is usually achieved by computing a mapping $\mathbf{F} : \mathbb{R}^d \rightarrow \mathbb{R}^q$ from data space to latent space with $q \ll d$ for a given set of n patterns (images) $\mathbf{y}_i \in \mathbb{R}^d$ with $i = 1, \ldots, n$. When dealing with imaging data, the problem of defining an appropriate measure of similarity arises. Typically, morphological features are extracted and a simple but well known metric such as the Euclidean distance is used (de la Calleja & Fuentes 2004; Wijesinghe et al. 2010).

In this paper, we present a novel approach based on a computationally intensive rotation and flipping invariant similarity measure. We employ a modified version of self-organizing maps (Kohonen 1989). Several non-linear dimensionality reduction techniques have been tested, and the approach from the area of artificial neural networks based on embedding patterns at fixed latent positions performed best. Our Parallelized rotation/flipping INvariant Kohonen map (PINK) framework makes use of multi-core CPU/GPU environments.This paper is structured as follows. In Section 2 we describe the PINK framework in general, the adopted similarity measure, and the parallelization steps. The data-set as well as the required pre-processing steps are described in Section 3 along with the design of the experiments. After presenting the empirical results in Section 4, we draw our conclusions in Section 5.

2. PINK

We promote a framework which generates a classification scheme in an unsupervised way and, thus, permits a semi-automatic analysis of the data by the user. In particular, we employ Kohonen-maps as a simple yet effective dimensionality reduction technique, which in our case, projects data to a two-dimensional map. Kohonen-maps are a specialized form of neural networks where every fixed node/neuron $\mathbf{p} \in P$ of the latent space contains a derived prototype after having trained the model. The neurons $P = \left\{\mathbf{p}_j = (\mathbf{w}_j, \mathbf{c}_j) \mid \mathbf{w}_j \in \mathbb{R}^d, \mathbf{c}_j \in \mathbb{N}^2, j = 1, \ldots, \mu_P\right\}$ map every prototype or weight-vector $\mathbf{w}_j$ to a coordinate $\mathbf{c}_j$ in the map. In the training phase, iteratively t times the n patterns $\mathbf{y}_i \in \mathbb{R}^d$ with $i = 1, \ldots, n$ are applied to the map. By calculating a similarity measure $\Delta(\mathbf{y}, \mathbf{w}_j)$ between a pattern $\mathbf{y} \in \mathbb{R}^d$ and the weight $\mathbf{w}_j \in \mathbb{R}^d$ of every node $\mathbf{p}_j \in P$ of the map, the closest (winning) neuron $q(\mathbf{y}) = \mathrm{argmin}_{j=1,\ldots,\mu_P} \Delta(\mathbf{y}, \mathbf{w}_j)$ is determined. Then the neurons are updated based on the distance to the winning neuron in the map $d(\mathbf{c}_{q(\mathbf{y})}, \mathbf{c}_j)$ and the number of applied iterations t via a training function $f(d(\mathbf{c}_{q(\mathbf{y})}, \mathbf{c}_j), t)$. This is done by setting the weight-vector $\mathbf{w}_j$ of a neuron $\mathbf{p}_j$ to the new value $\mathbf{w}'_j = \mathbf{w}_j + (\phi^{(j)}(\mathbf{y}) - \mathbf{w}_j) \cdot f(d(\mathbf{c}_{q(\mathbf{y})}, \mathbf{c}_j), t)$, where $\phi^{(j)}$ is the identity function in the standard Kohonen-map algorithm and will be used here to align the coordinate systems of $\mathbf{y}$ and $\mathbf{w}_j$. The function $f(\mathbf{c}_{q(\mathbf{y})}, \mathbf{p}_j), t)$ consists of a distance-dependent part and an iteration-dependent part. Currently PINK supports the use of a Gaussian or a Mexican hat as the distance component while using a simple linear damping based on the number of iterations t. Other distance components as well as iteration-dependent functions can be easily added.

The PINK framework allows to train Cartesian as well as hexagonal maps, both in a continuous repeating or edge-limited version. Because the hexagonal shape has six instead of just four distinct directions and just allows for natural numbers it was added to the framework. We could not determine huge differences when comparing the results based on astronomical images, besides that corner effects on the edge-limited version are not as dominant on the hexagonal map as on the Cartesian version. After the training phase is finished, one is able to match an image/pattern $\mathbf{y}$ to the derived prototypes P and thereby retrieve a coordinate c in the map. By inspecting and annotating the derived prototypes, a scientist inspects all matching objects at once. Therefore the amount of objects to be inspected is reduced to the number of prototypes in the map.

2.1. Similarity Measure

As described above, the training of the Kohonen-map depends on a similarity measure $\Delta(\mathbf{y}, \mathbf{w}_j)$ between the image y and the weight of the neuron $\mathbf{p}_j$. When inspecting images by eye, the brain automatically scales, aligns, distorts, and interpolates the information such that objects are perceived to be similar or not. Pre-processing the images to align them to the principal axis of their main component and using a simple pixel-wise Euclidean distance was one of the first approaches to deal with rotation. In the past, we carried out multiple tests with rotation invariant similarity measures (Polsterer et al. 2012). Up to now, we achieved the best results with Fourier transformed circular slices of the images. This method has the limitation of losing the information of complex and weak structures and therefore just allowed a sorting based on dominant morphological features. For the imaging data at hand, a rotation and flipping invariant similarity measure is essential to achieve satisfying results. To calculate the similarity, our approach basically calculates the Euclidean distances for all possible rotations/flipped/un-flipped objects in the map to determine the best match (see Figure 1). It can be shown that this operation still gives rise to a valid distance metric.

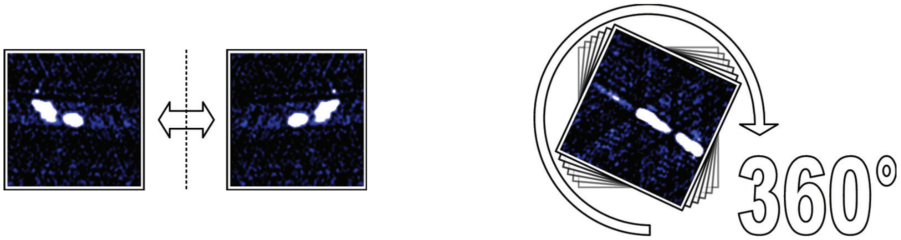

Figure 1. Both image transformations as they are applied to measure the similarity are shown exemplarily. The flipping (left) is shown on `FIRSTJ075843.0+611936` and the rotation (right) is shown on `FIRSTJ072529.5+614732`.

2.2. Speedup via Parallelization

Since the considered brute-force comparisons between an image $\mathbf{y}$ and all the neurons P are computationally very demanding, this task depicts an ideal candidate for massively parallel implementations. The current version makes use of modern multi-core systems while the GPU-based version is currently further optimized. In a first step all rotated/flipped versions of the image $\mathbf{y}$ are created in parallel by using a set

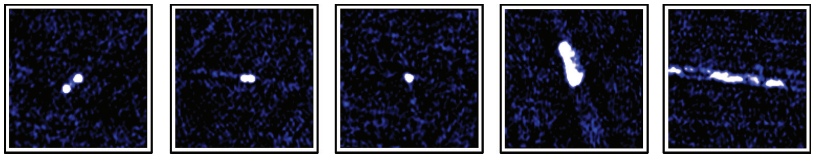

Figure 2. Examples for the imaging data used. Note the correlated noise in the background as well as the residuals from the imaga generation process. The objects shown are: `FIRSTJ070159.0+621123`, `FIRSTJ070153.7+640348`, `FIRSTJ070406.6+625235`, `FIRSTJ070557.2+625303`, `FIRSTJ072749.3+614904`.

of image transformations $\Phi = \{\phi_1, \ldots, \phi_N\}$. Next the Euclidean distance between the pre-transformed images and all neurons is calculated in parallel and the winning neuron $q(\mathbf{y})$ is determined. Finally, the weights $\mathbf{w}_j$ of all neurons are modified in parallel. In the update step, the alignment of the neurons with respect to the training image $\mathbf{y}$ must be considered (via a proper choice of $\phi^{(j)}$).

3. Data and Experiment

To test the performance and usability of our approach, we performed experiments on synthetic data as well as on real astronomical images. As the synthetical data was just use to ensure that we are able to reproduce simple geometric shapes, only the results on the real-word data will be shown.

The data used for our experiment is radio-synthetis data taken from the Radio GalaxyZoo project (`http://radio.galaxyzoo.org`). All of the 206,399 radio images from the FIRST survey (Becker et al. 1994) that are available have been processed with PINK. In Figure 2 some images of the used data are presented. We preprocessed the data in the following way: Regular cutouts with 128 px × 128 px have been created even though just 64 px × 64 px were used in the map. This size allows to create rotated versions for the similarity measure without having non-valid pixels in the corners. Note that just $\sqrt{2} \times 64$ px would be an appropriate size in our case. In addition, the images where scaled to values between 0 and 1 and every pixel below 2σ was masked as background and set to an according value. Finally, a hexagonal map of the size of 21 × 21 nodes was trained with the pre-processed data.

4. Results

After having performed the training on the 200 k images from Radio GalaxyZoo, we retrieved the map presented in Figure 3. This resulting map shows the derived prototypes which allow a clear separation into different morphological classes (Figure 4). Thus, by inspecting the map, one can basically analyze all objects represented by the prototypes simultaneously. For some objects, a heat map was created showing the regions in the map that match best (see Figure 6). Based on this mapping to the prototypes, it is possible to transfer the annotations created for the map directly to every individual image. Those objects that are not well represented by the prototypes can directly be extracted by using the absolute similarity value of the best match. In our experiment, just a fraction of a percent turned out to be such an outlier based on the analysis of the distribution (Figure 5) of the absolute similarity values. They can be concerned as

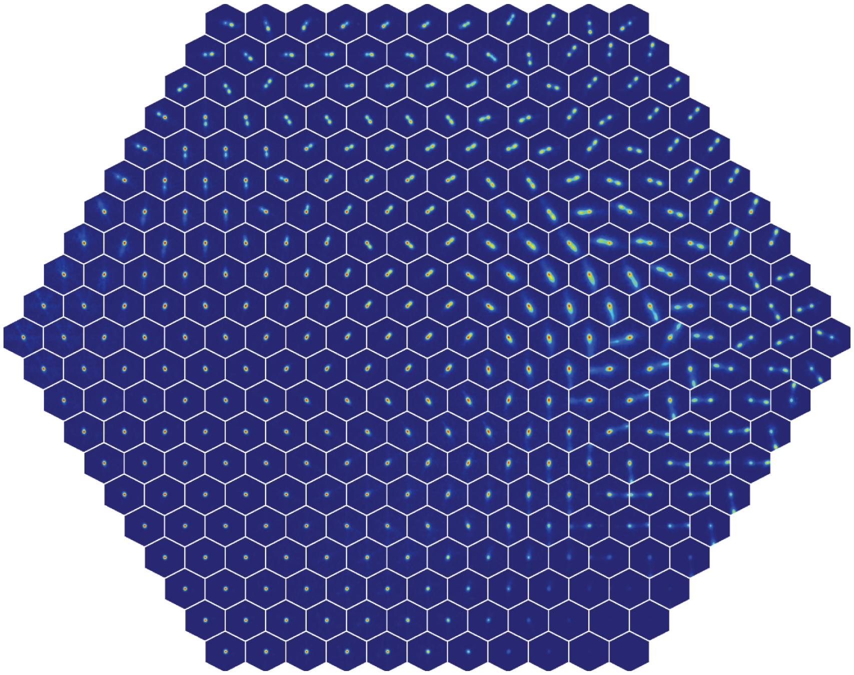

Figure 3. Resulting hexagonal Kohonen-map containing the derived prototypes.

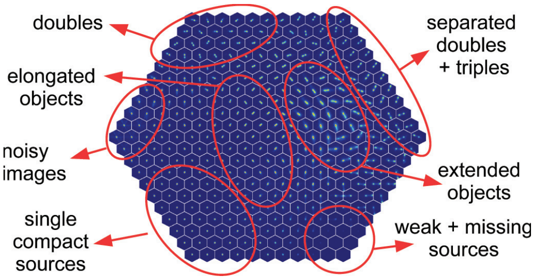

Figure 4. Resulting overview of the Kohonen-map with morphological classes being marked and labeled in red.

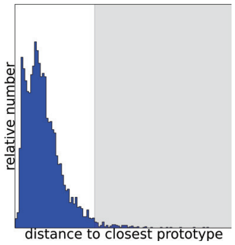

Figure 5. Distribution of the distance of the images to the prototypes. The area of the outliers is marked gray.

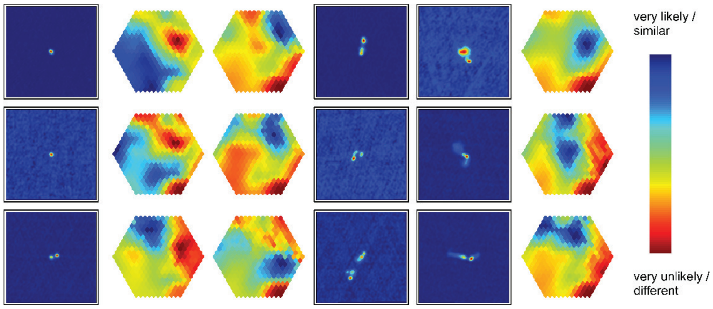

Figure 6. For some selected objects, the associated heatmaps are shown. Each individual heatmap directly reflects the similarity of an object to all the neurons in the map (blue represents similar while red reflects different neurons).

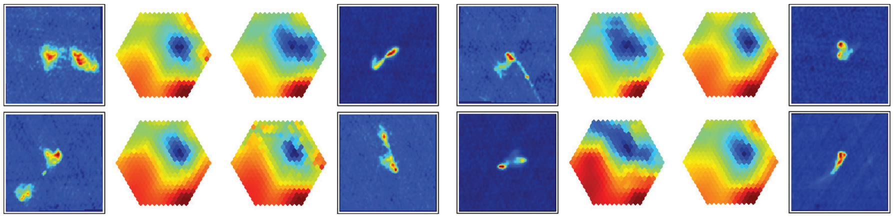

Figure 7. A list of outliers which have been selected based on the quality of their fit to the prototypes in the map. The corresponding heatmaps give an idea to which prototypes they could belong to even though they are not as good represented by the prototypes as the majority of objects.

interesting objects which require manual inspection by an expert. In Figure 7, some of the outliers that have been automatically extracted are shown. All the extracted outliers show interesting morphological features.

5. Conclusion

The proposed general method shows that unsupervised dimension reduction techniques can help astronomers to analyze huge amounts of data. Besides retrieving a classification scheme, one is able to efficiently detect outliers. Those are the objects which need to be analyzed by an expert. The majority of similar objects just needs to be inspected on the basis of a few representatives.

Acknowledgment

CI gratefully acknowledges support from The Danish Council for Independent Research through the project "Surveying the sky using machine learning".

References

Ahn, C. P., et al. 2014, ApJS, 211, 17. 1307.7735

Becker, R. H., White, R. L., & Helfand, D. J. 1994, in Astronomical Data Analysis Software and Systems III, edited by D. R. Crabtree, R. J. Hanisch, & J. Barnes, vol. 61 of Astronomical Society of the Pacific Conference Series, 165

de la Calleja, J., & Fuentes, O. 2004, MNRAS, 349, 87

Kohonen, T. 1989, Self-Organization and Associative Memory (Springer)

Kramer, O., Gieseke, F., & Polsterer, K. L. 2013, Expert Systems with Applications, 40, 2841 . URL http://www.sciencedirect.com/science/article/pii/S0957417412012432

Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R. C., Raddick, M. J., Szalay, A., Andreescu, D., Murray, P., & Vandenberg, J. 2011, MNRAS, 410, 166. 1007.3265

Polsterer, K. L., Gieseke, F., & Kramer, O. 2012, in Astronomical Data Analysis Software and Systems XXI, edited by P. Ballester, D. Egret, & N. P. F. Lorente, vol. 461 of Astronomical Society of the Pacific Conference Series, 561

Wijesinghe, D. B., Hopkins, A. M., Kelly, B. C., Welikala, N., & Connolly, A. J. 2010, MNRAS, 404, 2077. 1001.5322