Polish Academy of Sciences
Bartycka 18, 00-716 Warsaw, Poland
tel: +(4822) 841 00 41, +(4822) 3296 100
fax: +(4822) 841 00 46
email: camk@camk.edu.pl
http://www.camk.edu.pl
NICOLAUS COPERNICUS ASTRONOMICAL CENTER

Warsaw, October 28, 2024

Council of the Discipline of Physics

National Centre for Nuclear Research

"Search for galaxy mergers in big sky surveys"

doctoral thesis by Luis E. Suelves,

supervised by prof. Agnieszka Pollo and dr William J. Pearson

Referee's report

The PhD thesis by Luis E. Suelves presents the enormous progress made in data analysis concerning source identifications in the large sample of sky images provided by modern telescopes. The methods described are related to the galaxy mergers, but most probably they can be used to any type of photometric data, where the enhancement of flux is occurred due to the dynamical collisions. Of course, for each case of astrophysical sources training has to be made on different sets of parameters, but in general we observe the huge use of Machine Learning (ML) and Neural Networks (NN) techniques in modern astrophysics. It is worth to notice that in this year the Nobel Prize in Physics was awarded: “for foundational discoveries and inventions that enable machine learning with artificial neural networks”. And this is the subject of Luis Suevles’s thesis, where he describes, how ML and NN can be utilized to identify galaxy mergers, which represent an important stage of galaxy evolution. The dissertation presents a new and elegant way of merger’s identification from large sky surveys recently taken by Sloan Digital Sky Survey (SDSS) and Subaru/Hyper Suprime-Cam (HSC). Both surveys aim to collect the high-resolution galaxy images for the big sky areas, and Luis Suelves shows that ML and NN identifies mergers better than visual inspection made by human eye. Especially he demonstrated that the sky background error is a valid measurement to find mergers, and he used this method to several different largest datasets, implying several new conclusions about the meaning of the surroundings of galaxies and dithering effect. In all his work, the data analysis was made with the use of ML and NN training, if necessary.

The PhD thesis by Luis Suelves contains 7 chapters, where the first is an introduction on galaxy mergers and on the possible ways of finding them in the dens field of data. Second chapter introduces datasets used in the dissertation for different merger identification methods.

In the third chapter methodology of machine learning and photometric neural network is characterized and addressed to the exact dataset. It is worth to note that the each presented method is addressed to the specific dataset, depending on characterization of the telescope that was used during data collection. Next chapters 4, 5 and 6 describe results obtained by Luis Suelves, and either published in two papers or being prepared for future publication. The last chapter gives a summary of all achievements obtained by the author.

In the introduction chapter, main morphological features of galaxies and its mergers are specified, together with different ways of their identification. I was a little disappointed while reading this part, since I was expecting more information about machine learning given from the first principle. It was rather a history of using this method for different datasets instead of explanation how the method works or what do we need to make it working, beside the datasets of course. Then again long paragraphs are addressed to the structure of telescopes, especially those dedicated of CCD principles I found very useful. But I totally missed the discussion of what is the definition of extended source?

1) What is typical look of the galaxy on the CCD focal plane? Is it larger than PSF itself? In general, PSF should be more less known and specific for the instrument. Therefore, I would appreciate the figure, where author compares PSF shape to the shape of galaxy itself, or galaxy merger as well.

2) The same with dithering, which can be traced as a characteristic of the instrument.

The size of PSF was roughly given in Chapter 2, but it is not clear to me how big the merger should be itself to be classified as a galaxy/merger. Several magnitude models have been presented here, but I miss the figure comparing them to get intuition on their meaning. All datasets have been presented as well as the motivation of using them, but author is refereeing to previous paper to read the details of how the dataset has been created. I missed some details here, for instance:

1) What is the exact definition of unit Maggie.

The same while reading chapter 3, about methodology. I had many questions, the answers I could find in the refereed papers, and of course one may say this is not the field of my research, but doctoral thesis should be the source of knowledge for other researchers. My questions are:

1) What is the meaning of activation function and what is its shape?

2) Can variation in error normalization be illustrated somehow?

3) Is the flow chart of data reduction developed by the author or it is a standard one?

Chapter 4 is based on already published paper, in which Mr Luis E. Suelves is the first author. It demonstrates clearly, that the background error increases accuracy in finding mergers up to 92,36%. As, I have understood, Mr Luis Suelves has written all codes, and implemented algorithms, made PCA of outcoming results and their dependence on the normalization. My general question is:

1) Is skyErr the script written or modified by author?

2) What is the meaning of initial training rate of $5x10^{-5}$?

3) How this training rate affects results?

Chapter 5 and 6 are using different data sets Galaxy Zoo Data Release 1, and deeper Subaru HSC images. In both cases the modification of sky background method was required, but results look promising. My comments are:

1) Figs 6.1 and 6.2 use “bottom” left and right images, but actually they are “top”.

2) How the new data reduction in g-band differs from the standard one?

3) It was not surprised to my intuition that dithering does not affect low surface structure around galaxies, but the most important is S/N value. What is the critical value of S/N to make the whole analysis valuable?

4) Is there any predictable image to be shown what is expected by LSST?

All presented figures are very clear and emphasise the high quality of data analysis made by author. I appreciate all of them, but I have to admit, that it was hard for me to start any discussion with obtained results. I fully accept them and I’m waiting for further development which I hope it will be made by author. By reading this paper, I have got a feeling that Mr Luis Suelves likes very much the topic that he works on. So, I can only ask him a naive question:

1) How much faster, in computational time, presented methods are? Can author compare CPU usage, when different ML and NN methods for merger identification are used?

Summarizing, the PhD thesis of Mr Suelves contains novel analysis of high-resolution images from SDSS DR6, Galaxy Zoo Data Release 1, and HSC-NEP. The above analysis can be directly applied to the future data from upcoming Large Survey of Space and Time (LSST), which will be carried by the Vera Rubin Observatory. The author is a part of LSST collaboration. The thesis content fully proves that the author made deep effort to understand how ML and NN can help in merger’s identification, he understands all limitations of the method, and he is ready to show what can make the method even better in the merger’s identification. For me it is clear that Luis Suelves can work in this field on the highest possible level. Summarizing, the presented thesis fulfill all the formal and customary requirements presented by the law for a doctoral dissertation. Therefore, I request the Council of the Discipline of Physics, to proceed to further steps needed to award Mr Luis E. Suelves the PhD title.

With Best Regards

Agata Różańska

Prof. dr hab. Agata Różańska