QSVM For Big Data Classification Of Astronomical Events

Qiskit Research Hackathon 2021

by

Quantstellar team

Ahh data data everywhere..
Yet, speed and accuracy in classification so rare !

How to gauge, is it a galaxy, quasar or perhaps a star ??
Ughh so irritating, could the classical world have a new Avatar ?

But hasn't Moore's law reached its saturation already ?
Don't worry, Astronomy and Quantum Computing is the new Confetti !!

Well, just relax, run the QSVM, and chomp a biscuit ,
Lo and behold ! The magnanimous power of Qiskit !!

By Rita Abani

Acknowledgments

We, Quanstellar team, would like to thank Qiskit Hackathon organiser team for creating this valuable learning opportunity and IBM for providing us with the excellent infrastructure. We would like to thank Dr

Emille E. O. Ishida for your positive energy and expert knowledge that have helped us to refine our research and bringing forward an exciting further research idea. We would like to thank Mehdi Bozzore for your guidance in quantum computing. We would like to thank Hossein Afsharnia for connecting us to Fink team.

I. Introduction

Our goal is to promote Qiskit and QML applications in the exciting and emerging field of Astronomy and contribute to the astronomical community with this open-source project. Through this Qiskit Hackathon, we have got the opportunity to collaborate with Dr Emille E. O. Ishida, the lead researcher of the SNAD project and a core member of the Fink project. The Fink Project aims to prepare for one of the most ambitious astronomical projects, called Legacy Survey of Space and Time (LSST), to be launched in 2023. LSST aims to chart the universe on a broader and deeper scale than all previous surveys combined. It is projected to capture more than 20 Terabytes of raw data and produce more than 10 million alerts every night. Dr Ishida shared that one of the biggest challenges in astronomical research is having a limited amount of labelled dataset for training while enormous amounts of observations as test data due to the extremely high cost of data labelling and telescope complexity requirements. Classical machine learning algorithms usually fall short in classification tasks where training samples are limited. In [3], Belis et al. addressed this issue by showing that Quantum Machine Learning algorithms (QSVM) could tackle the challenge of limited training samples by outperforming classical machine learning algorithms even with a limited number of qubits available in current hardware. Therefore, in this project, we aim to explore the feasibility of implementing QSVM for the classification of astronomical data with a minimum amount of training samples. Our primary research questions are:

Research Question 1: Can a hybrid approach of classical and quantum machine learning algorithms outperform conventional machine learning algorithms?

Research Question 2: Can we achieve quantum speed up with this hybrid approach?

II. Methodology

1. General Approach

To answer our research questions, we implement a two-phase approach. The first phase is the classical phase, where we build the classical support vector machine (SVM) algorithm based on conventional machine learning techniques. We then apply the quantum SVM (QSVM) on the second phase - the quantum phase. We optimize the QSVM parameters using different circuit designs and feature mapping methods.

Classical SVM is enhanced with multiclass kernel tricks to achieve more complexity and handle non-linearity in the dataset. In the classical phase, we use the complete DR16 dataset of 100,000 observations instead of a smaller subset (200 - 400 observations) as in the quantum phase. This setup is designed to answer our first research question as well as to handle current platform constraints. In the Quantum phase, we implement a hybrid model of the multiclass quantum algorithm on a classical training dataset. We use the Qiskit machine learning QSVM package to train our core model and then hyper-tune the parameters such as circuit depth and entanglement structure to optimize the model. The implementation steps are as follows:

1. Load the dataset DR16 with selected features
2. Create a balanced subset of DR16 with training, validation and test set
3. Set the dimensionality and number of qubits required
4. Initialize the feature map to build QSVM
5. Create Quantum Instance and train the model with hyperparameter tuning

6. Evaluate the model prediction with multiclass classical SVM as the benchmark

We evaluate the two algorithms on two main tests: performance accuracy and quantum speedup. The performance accuracy-test evaluates the level of prediction accuracy running on the test set, commonly known in classical machine learning as the confusion matrix [6]. The quantum speedup test compares the runtime of a quantum algorithm with that of the classical computer, given by the equation speed_up = Q/C (where Q is quantum runtime and C is classical runtime) [5]. We run our jobs on several different platforms to compare the runtime and performance consistency, including IBM qasm simulators on a local machine, online IBM Q qasm simulator and IBM Q real device (ibmq_16_melbourne).

2. Data and Descriptive Statistics

In our project, the dataset we used is called the Sloan Digital Sky Survey Data Release (DR16). It is an astronomical dataset consisting of detailed three-dimensional maps of the Universe ever made, with multi-colour images of one-third of the sky and spectra for more than three million astronomical objects. It has 16 predictors to classify the data into three classes - Star, Galaxy and Quasar. A subsample of the dataset is attached here.

Our experiment on DR16 is the proof-of-concept for The Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC) that will be in the next phase. PLAsTiCC is an open data challenge launched by the Fink team to classify simulated astronomical time-series data in preparation for observations from the Large Synoptic Survey Telescope (LSST). PLAsTiCC dataset contains a large dataset to classify astronomical time series data into one of the fourteen classes. A subsample of the dataset is attached here.

Our team started with five members from all walks-of-life, working on the project proposal. After our proposal was selected, we immediately "met up" and detailed our implementation plan. While model-driven tasks are divided amongst the programmers, literature research and documenting tasks are based on the whole team's efforts. We even have a brilliant poet in our team who gifted us with two great poems mentioned in this report.

3. Results

Throughout the project, we tried to find the answers to our research questions. Our extensive set of experiments showed that a hybrid approach of classical feature transformation and QSVM has outperformed classical SVM in terms of accuracy of multiclass classification. In Table 1. A comparison of classification accuracy and runtime of different combinations of our hybrid approach has been reported. According to the results, our hybrid approach of classical feature selection and transformation and QSVM has outperformed the classical SVM by a significant margin. However, in terms of runtime, classical SVM is a lot faster than QSVM.

Table 1. Performance Result

	No. of training samples	No. of test samples	Feature Selection and Transformation	Dimensionality Reduction	Feature map	Success Ratio (accuracy)	Runtime (s)
QSVM	100	150	Yes	Yes (ISO Map)	ZZFeatureMap	0.55	3584.3
SVM	100	150	Yes	Yes (ISO Map)	Rbf_kernel	0.5	0.026

QSVM	100	150	Yes	Yes (LLE)	PauliFeatureMap (Z, ZZ)	0.39	4124.3
SVM	100	150	Yes	Yes (LLE)	Rbf_kernel	0.33	0.06
QSVM	200	150	Yes	No	PauliFeatureMap (Z, X, ZY)	**0.81**	**6978.1**
SVM	200	150	Yes	No	Rbf_kernel	**0.63**	**0.04**

III. **Conclusion**

We presented our findings to Dr. Ishida and received a lot of positive feedback. Two main reviews are 1. Our results suggest a promising method to tackle one of the biggest challenges in big data astronomical research 2. This research is a solid first step to kick start the collaboration between the Quanstellar and Fink team, by applying our work on the PlasTiCC dataset as preparation for the 2023 LSST run. Moreover, we put extra effort into cleaning up our code repository with the hope to inspire future researchers to further expand this uncharted quantum territory.

Some improvements to our work can be made in the future, such as reducing the dimension on the data matrix and its complexity; optimising the QSVM quantum circuit, which can be achieved through a deeper investigation of the QSVM algorithm and customised feature map. From the technical point of view, we want to finish training our model on other real IBM quantum computer and increase the amount of test data for further performance robustness checks.

REFERENCES

[1] Montanaro, A. Quantum algorithms: an overview. npj Quantum Inf 2, 15023 (2016). https://doi.org/10.1038/npjqi.2015.23

[2] Lee, Joong-Sung, et al. "Experimental demonstration of quantum learning speedup with classical input data." Physical Review A 99.1 (2019): 012313.

[3] Belis, Vasileios, et al. "Higgs analysis with quantum classifiers." arXiv preprint arXiv:2104.07692 (2021).

[4]Ding, Chen, Tian-Yi Bao, and He-Liang Huang. "Quantum-inspired support vector machine." *arXiv preprint arXiv:1906.08902* (2019).

[5] Rønnow, Troels F., et al. "Defining and detecting quantum speedup." *science* 345.6195 (2014): 420-424.

[6] Powers, David MW. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." *arXiv preprint arXiv:2010.16061* (2020).