MACHINE LEARNING PROJECT
5^{th} SEMESTER
PES UNIVERSITY

CLASSIFCATION OF STARS AND QUASARS USING SVM [*]

[1]Vishwas N.S, PES1201701321
[2]R.Ananth, PES1201700286
[3]Achintya Shivam, PES1201700151

November 27, 2019

1 ABSTRACT

A problem that lends itself to the application of machine learning is classifying matched sources in the Galex (Galaxy Evolution Explorer) and SDSS (Sloan Digital Sky Survey) catalogs into stars and quasars.The problem is daunting due to the fact that there is no entity that separates stars and quasars.We have tried to develop a support vector machine model.To evaluate the efficiency of the classifier,the classifier gives us an accuracy greater than 90 for the catalogs given to us. Although spectrometric redshift data is itself a differentiator between stars and quasars, we have not used it in training,but have used it for cross-verification of the predictions. We have tried boosting to improve the performance of our model.We have also implemented ADABOOST, a boosting algorithm that helps us combine multiple “weak classifiers” into a single “strong classifier”

2 DATA AND PROBLEM STATEMENT

- A STAR is a type of astronomical object consisting of a luminous spheroid of plasma held together by its own gravity. The nearest star to Earth is the Sun.

[*] Github link: https://github.com/achintyashivam11/ML-CLASS-PROJECT.git

- A QUASAR, also known as quasi-stellar object, is an extremely luminous active galactic nucleus (AGN). The power radiated by quasars is enormous. The most powerful quasars have luminosities exceeding 1041 watts, thousands of times greater than an ordinary large galaxy such as the Milky Way.
- Our goal was to accurately predict the stars and quasars from the given datasets.The datasets given to us go by the names of catalog1,catalog2,catalog3,catalog4.
- Datapoints are considered from either the North-Galactic Region or the Equatorial Region .

3 IMPLEMENTATION

The libraries we have used in our project are numpy,pandas and cvxopt.We have also used scikit learn for splitting the dataset into training and testing data.We have trained our data on 0.8 fraction of the whole dataset and 0.2 fraction kept as testing data.Thus,an 80-20 model of training and testing which prevents overfitting.Using the cvxopt library,we have found out the lagrangian multipliers.The lagrangian multiplier then helps us in estimating weights and intercepts.We then have made a predict function which gives a fairly good accuracy on estimating stars and quasars.For predicting the accuracy,we have again used scikit learn.The steps followed for our project is as follows:

- Filtering out unwanted features from our dataset which are not required for our classification
- Also filtering out the spectrometric redshift from our training data so as that our model learns with our features as well
- applying our svm classifier
- svm classifier made with incorporating cvxopt for convex optimisation
- Lagrangian multipliers are found out using results from cvxopt
- Subsequently filtering out non-zero lagrangian multipliers
- We get an estimate of intercept and weight vectors once we get our lagrangian multipliers
- Then we have applied AdaBoost using our own svm classifier created as base classifier
- The weak learners in AdaBoost are SVMs with RBF kernels with a low gamma value and a low penalty value.
- AdaBoost works by putting more weight on difficult to classify instances and less on those already handled well. .

4 RESULTS

The following table consists the results for running different kernels on the different datasets that were provided to us. Results for catalog 4 are not present below since our implementation was not scalable to big datasets.

The ovarall accuracy (A), class-wise precision (P), recall (R) and F1-Scores (F1) have been reported for each of the catalogs and for each of the kernels implemented.

All of the below mentioned metrics are averages calculated using K-Fold cross-validation, with K being equal to 5.

		Linear Kernel				Gaussian Kernel				Laplace RBF Kernel			
Catalog	Class	A (%)	P	R	F1	A (%)	P	R	F1	A (%)	P	R	F1
Catalog 1	-1	95.99	0.75	0.7	0.72	95.83	0.97	0.52	0.67	96.45	0.93	0.61	0.72
	1		0.98	0.98	0.98		0.96	0.99	0.98		0.97	0.99	0.98
Catalog 2	-1	92.81	0.70	0.61	0.65	93.06	0.88	0.43	0.57	95.01	0.82	0.67	0.73
	1		0.95	0.97	0.96		0.93	0.99	0.96		0.96	0.98	0.97
Catalog 3	-1	92.99	0.70	0.61	0.65	95.62	0.84	0.74	0.78	95.62	0.86	0.70	0.77
	1		0.95	0.97	0.96		0.97	0.98	0.98		0.96	0.99	0.98

Table 1: Results calculated for three different kernels used in SVM

Adaboost with SVM					
Catalog	Class	A (%)	P	R	F1
Catalog 1	-1	96.76	0.96	0.63	0.75
	1		0.97	0.99	0.98
Catalog 2	-1	94.54	0.79	0.68	0.72
	1		0.96	0.98	0.97
Catalog 3	-1	95.27	0.84	0.69	0.76
	1		0.96	0.98	0.97

Table 2: Results calculated using Adaboost with SVM as the base classifier - Laplace RBF Kernel

5 CONCLUSION

The following is a summarised conclusion of our project:

- We were supposed to classify stars and quasars on the GALEX and SDSS data given to us.
- We trained our data on svm classifier and also on svm+adaboost classifier .

- SVM and SVM+Adaboost was implemented on our own with the support of online references and research papers mentioned in the reference section.
- Our model performed fairly good on the most of the datasets given to us .
- Although the SVM+Adabooost technique was used to improve the accuracy, the classifier performed well only on catalog 1 when compared to the normal SVM, and less accurate on catalog 2 and 3.

6 REFERENCES

1] Machine Learning in Astronomy: A Case Study in Quasar-Star Classification
Mohammed Viquar , Suryoday Basak , Ariruna Dasgupta , Surbhi Agrawal , Snehanshu Saha

2] Website : http://astrirg.org/projects.html

3] Website : https://pythonprogramming.net/