BYTE PANACHE COMPETITION

Foundations of Modern Machine Learning

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

HUB.DATA IIIT HYDERABAD

OBJECTIVES

The motive of this work is to produce and analyze the results obtained via SVM(support vector machine) classifier to understand how SVM classifier works on different data sets.

MATERIALS & METHODS

We have used three different datasets to analyse SVM classifier. We have acquired results for four kernel based settings i.e., RBF kernel, linear kernel, polynomial kernel and sigmoid kernel whose results have been depicted in the Results 1 and 2 sections, in detail. The three datasets utilized to obtain the results are:
1) **Pulsar Star Dataset**
2) **Diabetes Dataset**
3) **Breast Cancer Dataset**
These datasets have been acquired from kaggle and UCI repositories respectively.
–Pulsar Star dataset contains of 17,897 samples with 9 attributes.
–Diabetes Dataset contains 768 samples with 9 attributes.
–Breast cancer dataset contains 569 samples with 31 attributes.

REFERENCES

References

[1] Muller, Andreas C., and Sarah Guido. 2017. Introduction to machine learning with Python: a guide for data scientists.

[2] https://en.wikipedia.org/wiki/Support-vector_machine

[3] www.kaggle.com

[4] Sarkar, Tushar. "XBNet : An Extremely Boosted Neural Network." ArXiv abs/2106.05239 (2021): n. pag.

INTRODUCTION

SVM classification is one of the classification methodologies under supervised learning in machine learning. It is considered to be one of the efficient algorithms in existing classification methods in machine learning. Hence we are worked to **analyze and interpret** how SVM classifier works on **multiple different data sets.**

RESULTS 1

Considering the results obtained on the three different datasets on four different kernels under 3 different settings, we could interpret the following points:
–**Linear** Kernel is considered to be one of the most effective kernels compared to other three different kernels. The amount of learning using linear kernel is comparitively high than rbf, polynomial and sigmoid kernels.
–**Sigmoid** kernel was considered to be less effective than the other 3 kernels. The accuracy obtained via this kernel seemed to depreciated than others.
–The amount of learning via **Polynomial** kernel for Pulsar star and Cancer datasets was relatively high and could interpret the reason to be having more samples and more number of attributes respectively for the datasets.
–Similar to that of Polynomial kernel, the performance of RBF kernel seemed to be equal by maintaining the same trend of performing well on pulsar star and cancer datasets.
The obtained results have been tabulated and depicted in the Results 2 section along with the visual interpretation of the utilized datasets for svm classifier (kernel = RBF).
Note - All the datasets have been scaled before the experimentation.

FUTURE RESEARCH

Application of various mathematical approaches in terms of normalization and embedding SVM into Neural networks would bring out better results. One of such recent research works was supposed to be XBNet.

RESULTS 2

Datasets \ Kernel Type	RBF			Linear			Polymial			Sigmoid			Classification Accuracy
	C=1	C=100	C=1000	C=1	C=100	C=1000	C=1	C=100	C=1000	C=1	C=100	C=1000	
Pulsar Star	97.99%	97.99%	97.82%	97.99%	98.07%	98.07%	97.93%	97.93%	97.93%	88.58%	88.60%	88.60%	97.99%
Diabetes	81.97%	74.03%	71.43%	80.52%	80.52%	80.52%	74.03%	75.32%	75.97%	75.97%	67.53%	71.43%	80.52%
Cancer	98.25%	94.74%	94.74%	98.25%	95.61%	96.49%	90.35%	97.37%	94.74%	91.23%	88.60%	88.60%	98.25%

Diabetes   Cancer   Pulsar Star
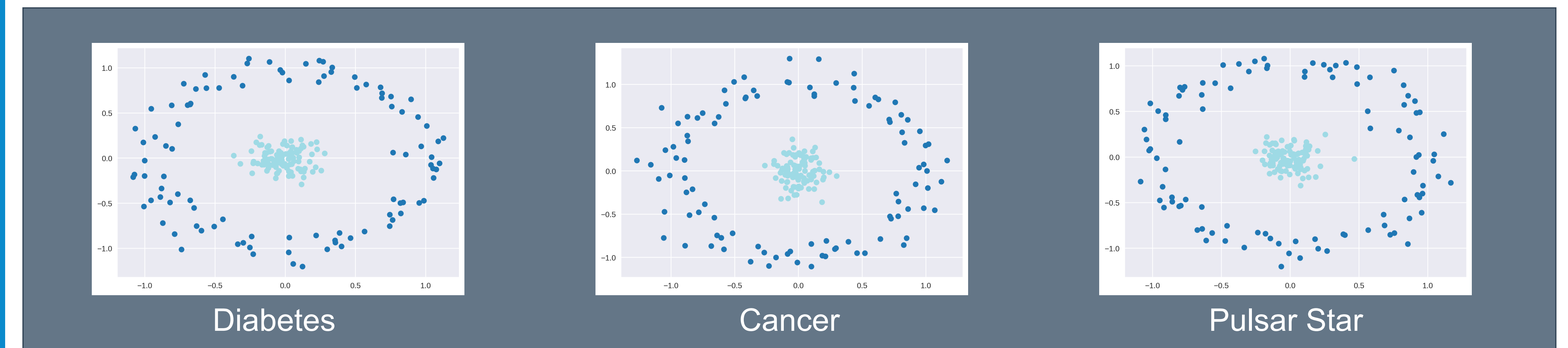
–The above illustrated table is regarding the obtained results. The description with respect to the analysis have been provided in the Results 1 section. Also, we have visualized the SVM classifier with respect to the 3 different datasets in a non-linear plots. These plots provide visual interpret-ability of how the data looks when plotted with respect to their contradictory samples.
–The classification accuracies obtained on each dataset have been also tabulated in the above table.The accuracy obtained on cancer dataset is more higher than compared to other to datasets, even though the results obtained via Pulsar star dataset is having a very minute marginal difference. The reason for this is because of number of attributes are comparatively higher by 22 units in this dataset. The pulsar star dataset is considered to have more number of samples and hence the accuracy is having a marginal difference compared to that with highest accuracy.

CONCLUSION

–From the detailed experimentation, we could understand that the dataset size plays a major role in classifying some specific entity. The number of samples and number of attributes play an equal role in classifying a specific object into a particular class.
–By this we can conclude that the dataset which is being used to classify a particular object must contain ample amount of attributes along with its samples, according to the observation, samples play a bit more vital role in classifying the data into classes.

CONTACT INFORMATION

Group A
1) **FMML20211000** - Gnyanesh Bangaru - fmml20211000@ihub-data.iiit.ac.in
2) **FMML20211001** - Dhruva Agarwal - fmml20211001@ihub-data.iiit.ac.in
3) **FMML20211002** - Anvitha Moila - fmml20211002@ihub-data.iiit.ac.in