Evaluation of neural network based classification methods for Gamma rays identification

Shalini Tyagi **Aishwarya Chandra Shekhar**

Abstract

Evolution of neural network methods has significantly contributed towards the accelerated development of the machine learning algorithms. While the objective and general concept behind these methods may have some gross similarity, they differ a lot with their implementation in terms of efficiency and accuracy of the results. This paper aims to present a critical evaluation of two neural network algorithm models, performed on a data set collected from MAGIC, a Cherenkov telescope experiment. The task is to classify gamma signals from a data set containing a combination of gamma and hadron signals. We compared two algorithms for this study: Multilayer Perceptron with backpropagation (MLP) and Support Vector Machines (SVM). The best model is selected by varying the hyperparameters for both the models in a grid search manner and then validated through 10-fold cross validation. The tested results from the best evaluated models are compared by confusion matrices and Receiver Operating Characteristic curves (ROC). Through our experiments, it was found that for such a classification problem, SVM model performed much better than the MLP algorithm.

1. Introduction:

MAGIC (**Major Atmospheric Gamma Imaging Cherenkov Telescopes**) is a system of two Imaging Atmospheric Cherenkov telescopes situated at the Roque de los Muchachos Observatory on La Palma, one of the Canary Islands, at about 2200 m above sea level. MAGIC detects particle showers released by gamma rays, from the outer universe, using the Cherenkov radiation, i.e., faint light radiated by the charged particles in the showers. It is designed to provide vital information on several established gamma-ray sources, such as Active Galactic Nuclei, Supernova Remnants, Gamma Ray Bursts and Pulsars.

Due to atmospheric radiations, the ground based telescope tends to collect other overwhelming events of hadrons, also called background. In order to understand the gamma ray sources, it is important to separate gammas from other particles. There is only a weak discrimination between the gamma and background events, making the data an excellent proving ground for the classification techniques. To separate the gamma signals from the overwhelming background signals would be a daunting task, as both have very similar characteristics. Classification of signals plays a vital role for astronomical analysis of gamma ray objects, and any small improvement in the classification accuracy would be significant in these analysis tasks [1].

1.1. Multilayer Perceptron (MLP) : Multilayer perceptron is a type of deep artificial neural network which is composed of more than one perceptron. This network consists of an input layer to receive the signal and output layer to produce the outcome about the inputs. Generally, MLP are used for supervised learning tasks. Which means there is a training set which consist of input and output values and the network must learn the dependency between input values and output values.

MLP with backpropagation algorithm consists of two steps: Forward pass and the backward pass. During the forward pass all weights do not change and the input are multiplied by the weights for each hidden neurons. In backward pass, weights are gone backward to change the weights recursively for improving the accuracy. MLP has a few advantages: 1. they are good to model with non-linear data with large number of inputs 2. Once the model is trained the predictions are very fast. However it is very computationally intensive and time consuming to train.

1.2. Support Vector Machines (SVM):

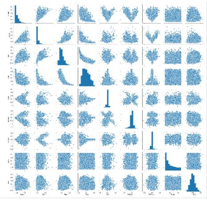
Support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze the data used for classification or regression analysis. Given a set of training samples, each sample is marked as belonging to one or the other of two categories. The SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The best hyperplane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points. If the data points are not linearly separable, SVM can introduce an additional dimension, using its 'kernel function' hyperparameter. By doing so, the data points are distributed in a way that they become separable by a hyperplane. SVM has certain advantages over MLP: 1) The kernel trick: It can use any number of hyper dimensions to transform the data into another dimensional space so that they can be separated using a hyperplane.2) It uses few data points for calculation, only the support vectors, hence reducing the computation time.

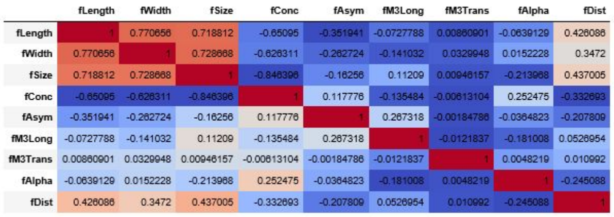
2. Dataset: For the purpose of the coursework, 'MAGIC Gamma Telescope Data Set' from the UCI repository has been chosen to compare the performance of the two models [2]. This dataset has 19,020 observations with 10 variables and 1 class. Additionally, the dataset was not in scale so the normalization is performed to bring the dataset in scale between the range [-1,1]. This data set is not balanced because class gamma 'g' has more number of observations than the class hadron 'h'. So we use oversampling approach in which the minority class is over-sampled by generating 'synthetic' observations[3]. After using SMOTE, this data has about 24286 observations and now data is scaled and balanced. Also, class has been changed from categorical values to numerical such as value for 'g' is 1 and for 'h' is 2 in the dataset.

2.1. Initial Data Analysis: The distribution of different features is shown in Figure 1 below:

- The data set contained various numeric measurements (in counts, mm, degrees and ratios) and there were no missing values.
- Significant differences between classes are in the mean and standard deviation of the ellipse length ('fLength').
- The mean of the angle between the main axis of the image and the center of the elliptical cluster ('fAlpha') is notably different between classes - this can be attributed to the distribution of the data, which is positive skewed for gamma and more uniform for hadron.
- The distribution of the distance between the centre of the camera and the centre of the shower image ('fDist') appears to be Gaussian, which can be observed visually
- 'fConc' and 'fConc1' were highly correlated. In both features, there wasn't a significant difference between the individual class means and standard deviations, however, the difference in skewness was greater for 'fConc'. Hence, 'fConc1' was identified as the redundant feature and removed from the dataset.

Thus, our final data set consists of 9 (nine) Predictor Variables and 1 (one) Target Class.





	fLength	fWidth	fSize	fConc	fAsym	fM3Long	fM3Trans	fAlpha	fDist
fLength	1	0.770656	0.718812	-0.65095	-0.351941	-0.0727788	0.00860901	-0.0639129	0.426086
fWidth	0.770656	1	0.728668	-0.626311	-0.262724	-0.141032	0.0329948	0.0152228	0.3472
fSize	0.718812	0.728668	1	-0.846396	-0.16256	0.11209	0.00946157	-0.213968	0.437005
fConc	-0.65095	-0.626311	-0.846396	1	0.117776	-0.135484	-0.00613104	0.252475	-0.332693
fAsym	-0.351941	-0.262724	-0.16256	0.117776	1	0.267318	-0.00184786	-0.0364823	-0.207809
fM3Long	-0.0727788	-0.141032	0.11209	-0.135484	0.267318	1	-0.0121837	-0.181008	0.0526954
fM3Trans	0.00860901	0.0329948	0.00946157	-0.00613104	-0.00184786	-0.0121837	1	0.0048219	0.010992
fAlpha	-0.0639129	0.0152228	-0.213968	0.252475	-0.0364823	-0.181008	0.0048219	1	-0.245088
fDist	0.426086	0.3472	0.437005	-0.332693	-0.207809	0.0526954	0.010992	-0.245088	1

Figures 1 & 2: Features Distribution and Heat Map

2.2 Hypothesis statement:

According to Praveen Boinee et.al [1], MLP had a better accuracy than the SVM. So our Null (Ho) and Alternative (H_1) Hypothesis are as follows:

Ho: MLP will not perform better than SVM
H_1: MLP will perform better than SVM

3. Methods:

In this section, we describe how we divided our main data into training, validation and testing data sets. We also describe the architecture and hyperparameters used for building the MLP and SVM models.

3.1. Methodology:

The main data set was divided into 70% training and validation data and 30% test data, using Randperm function. (The testing data was used as common for the testing of both MLP and SVM models). The training and validation data was further divided into 70% training and 30% validation data.

The best model was selected by conducting a grid search on the hyperparameters of both MLP and SVM models. The separation of data into training and validation data was done using the 10-fold cross validations, to prevent overfitting. To select the best hyperparameters, grid search was conducted on the training and validation data with different values of hyperparameters. The hyperparameters that had the lowest average loss, over 10-folds of cross-validations, were selected as the hyperparameters for the best model.

For the algorithm comparison, each of the 2 models, with selected best hyperparameters, were re-trained and tested on the common test data. Since in this experiment, the aim is to identify the gamma rays (the positive class), the algorithm that had the higher measures of accuracy, F-measure, sensitivity and precision was chosen as the better algorithm. For the MLP, part of the training data, validation data was used for early stopping. So that the test data was not fed during the training process. When the best parameters have been chosen then the best model was tested on the test data.

3.2. Architecture and Parameters used for the MLP:

Traindgx (A ‘Gradient Descent’ with momentum and adaptive learning rate backpropagation’) is a training function which updates the weight and bias according to adaptive learning rate and gradient descent momentum. The learning rate is a hyperparameter which helps to control the weight according to the gradient descent which means how much do be need to change the weight. The momentum is the term which is used to speed the network training. It adds the percentage of the prior weight changes to the current weight changes[4]. It is useful to have two outputs for the classification problem in the neural network.

Multi layer perceptron with back propagation algorithm is initialized with the weights randomly during each training process. As a result, the accuracy of the test data may be different. For training the neural network, we set a number of hidden layer, momentum and learning rate. Additionally to avoid the overfitting, we used K-Fold cross validation function on the training data and trained the model efficiently. For MLP, one hot encoding has been used to convert the categorical variable in to a form that helped the neural network to perform well in prediction.

3.3. Architecture and Parameters used for the SVM:

The hyperparameters chosen for SVM were Kernel function and Box Constraint.

Kernel methods are a class of algorithms for pattern analysis, whose task is to find and study general types of relations in datasets. If the raw data points are distributed in a manner that they are not separable using a hyperplane, then kernel method is applied. This method transforms the data into another dimension, such that the data points become separable by a hyperplane and hence assist in better classification of the classes present in the data set. The Kernel Functions tested for SVM were ‘Linear’, ‘Gaussian’ and ‘Polynomial’. For Polynomial function, polynomial orders tested were ‘2’, ‘4’ and ‘5’.

Box Constraint is described as a parameter that penalizes vectors laying outside the epsilon-margin. A higher box constraint causes higher costs. Increasing the box constraint imposes a hard margin and does not allow data points in the margin, leading to fewer support vectors. However, increasing the box constraint can lead to longer training times [5]. The Box Constraints tested were ‘4’, ‘5’ and ‘6’.

4. Results, Findings & Evaluation

4.1. Choice of parameters and experimental results

To select the best model, different models were run with different values for hyperparameters, using Grid Search. To reduce overfitting and make the model more generalised, 10-fold cross validation was done on training and validation test during the Grid Search.

For MLP models, it was noticed the performances of the model were not significantly changing when the values of the parametre was very small. For MLP, the values of error were varying according to the all hyperparameters changed. The range of parameters such as learning rate and momentum was in range [.1 to 1]. Also the errors of different models between 0.376 and 0.230.The best model was chosen based on the best parameters values (**learning rate = 0.3, momentum = 0.9 and hidden layer = 10**).

For SVM models, it was seen that the model performances were most sensitive to the changes in the Kernel function, with Linear function having the highest losses for all values of Box Constraint. The losses of different models ranged from 0.182 to 0.282. The higher order Polynomial functions performed better than other models for different values of Box Constraints. This means that the data points are not linearly separable and need to be projected on additional dimensions to be able to be separable by a hyperplane. The best model had **Polynomial kernel of order 4 and Box Constraint of 6**. The values of losses for different combinations of hyperparameters are shown in the chart below:

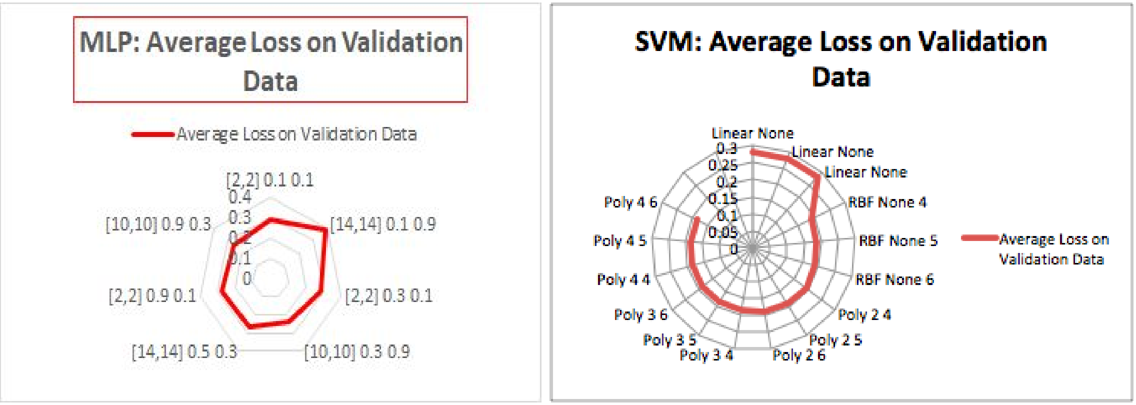
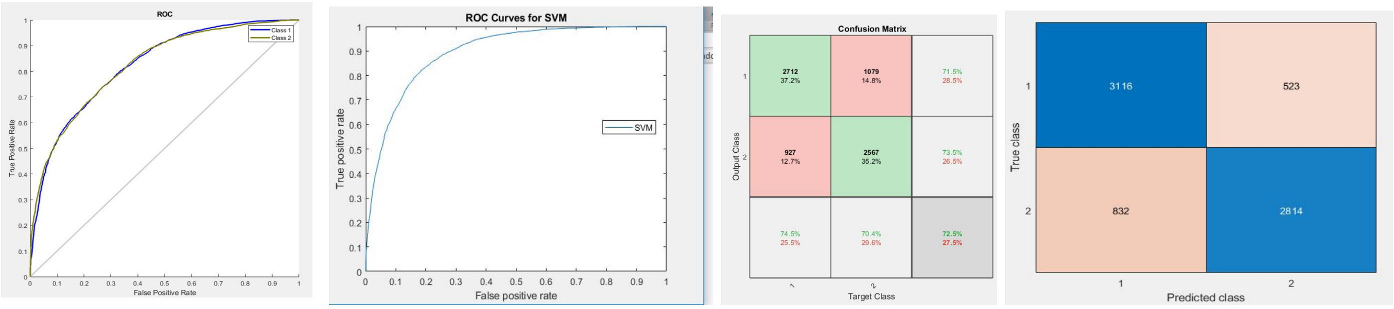


Figure 3: Experimental results for calculating loss with different Hyperparameters



Figures 4 & 5: Confusion Matrices and ROC curves (Left:MLP, Right:SVM)

4.2. Analysis and Critical Evaluation

In this section, the performance of MLP and SVM models are compared on the basis of the values of their Confusion Matrices. The best models selected for each algorithm (as discussed above) are re-trained and tested on the common test data and corresponding confusion matrices are plotted. SVM misclassified about 30.15% less data than MLP on Validation data set

and about 32% less data on Test data set. This shows that SVM performs much better than MLP on unseen data. The performance measures for SVM and MLP are as shown below:

	MLP	SVM
Accuracy	72.46%	81.40%
F-measure	73.00%	82.14%
Sensitivity	71.50%	85.62%
Precision	74.50%	78.92%
Time(sec)	4.307	9.894

Table 1: Performance measures for MLP and SVM models

From the table, it can be seen that the accuracy of SVM on Test data is about 12.34% more than that of MLP. Sensitivity (Recall) is a measure that shows whether all the positive classes (gamma rays, here) are retrieved. And Precision gives a measure of the relevance of retrieved positive classes. From the table, it can be seen that SVM retrieved about 19.74% more of relevant data than MLP. Also, relevance of the data retrieved by SVM was about 5.93% more than that of MLP. F-measure, which is the weighted average of Recall and Sensitivity is also higher for SVM. All these measures show that for this particular data, SVM classified the correct Gamma rays much better than MLP.

From figure 4 above, it can be seen that Area Under Curve for SVM is bigger than that for MLP. The AUC for SVM is 0.897. This shows high ability of SVM in accurately classifying the binary classes.

The higher polynomial order used and the higher box constraint employed in SVM, helped SVM in classifying Gamma rays better than MLP. However, the time taken by SVM to classify the same amount of data is almost double of that taken by MLP. Thus, the greater accuracy of SVM is at the cost of time taken for classification.

From the data distribution curve above in figure 1, it can be seen that the distribution of features in the data set are Linear, Gaussian as well as Polynomial. However, MLP, with gradient descent, contains a Linear activation function for output layer [6]. Thus, it classifies all the features linearly. On the other hand, SVM has Kernel trick, with which it can introduce any number of hyper-dimensions to make the data separable by a hyperplane. From the observations above, it can be seen that best SVM model uses Polynomial kernel of order 4. This kernel trick of SVM helps in better classification of features, in comparison to MLP.

5. Conclusion

Our study compared the performance of two trained models, Multilayer Perceptron and Support Vector Machine, in classifying Gamma rays from an unseen data set containing combination of Gamma and Hadron rays.

From our Hypothesis statements in section 2.2, we tried to test whether MLP will perform better than SVM in classifying the Gamma rays. However, our results (performance measures as shown section 4) show that SVM performs much better than MLP on both seen and unseen data. Thus, we **Do Not Reject our Null Hypothesis.**

We learnt that the comparatively poor performance of MLP could be attributed to the fact that MLPs are highly contingent on its neuron`s weights, resulting in significant variations in the accuracies.

5.1 Future Work

1. In future, we could work on Ensemble Classifiers that combine 'base classifiers' to increase the classification accuracy [1]. Ensembling techniques, such as AdaBoost and Bagging could be used for this purpose.
2. In this study, SMOTE was applied on the entire data set. In future, we can try to apply SMOTE only on the Training data and not on the Test dat. This would help test the accuracy of models on the unseen imbalanced data set[7].

6. References

[1] Praveen Boinee.et.al(2007) Ensembling Classifiers – An Application to Image Data Classification from Cherenkov Telescope Experiment ,World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:1, No:12

[2] R.K. Bock, P. Savicky (2007) UCI Machine Learning Repository[http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[3] Chawla.N.V., Bowyer.K.W., Hall.L.O., Kegelmeyer.W.P(2002) 'SMOTE: Synthetic Minority Over-sampling Technique' *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002

[4] Nii O. Attoh-Okine (1999) ' Analysis of learning rate and momentum term in backpropagation neural network algorithm trained to predict pavement performance ' Department of Civil and Environmental Engineering, Florida International University:Miami, USA

[5] Andrew, A.M., 2000. An Introduction to Support Vector Machines and Other KernelBased Learning Methods by Nello Christianini and John Shawe-Taylor, Cambridge
University Press, Cambridge, 2000, xiii+ 189 pp., ISBN 0-521-78019-5 (Hbk,£ 27.50).

[6] MATLAB® Programs for Neural Systems (2013) Available at: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118534823.app4
(Accessed on 28/03/19)

[7] Blagus.R and Lusa.L (2013) SMOTE for high dimensional class imbalanced data