A comparative study on various classifier models for prediction of celestial objects

Dr. Anwesa Sarkar

Assistant Professor of Physics

Sundarban Mahavidyalaya,

Kakdwip, South 24 Parganas, 743347, India

Abstract: *Classification of celestial objects like stars, galaxies and quasars is one of the most challenging and* fundamental problem in astronomy. Continuous accumulation of large volume data due to technological advancement of telescopes and observatories requires automation of classification. Various machine learning methods have been deployed now a days to do the classification accurately. In *this paper, comparison among the performances of different types of classifiers have been discussed. XGBoost appears to be the most efficient one.*

Keywords: **classification, stars, quasars, decision tree, XGBoost**

1. INTRODUCTION

Now a days various machine learning techniques grab the attention of the physicists due to its mammoth ability to extract insight from large volume of data. Due to technological advancements of orbital telescopes and large ground based observatories, huge volume of complex astronomical data containing high quality information are accumulated continuously. The availability of such a high volume data has associated flip side too; as the data becomes large, the difficulty to analyze the same increases manifolds. Machine learning techniques are rigorously used in observational astronomy to clean, mine and classify the data.

Classification of astronomical objects like stars, galaxies and quasars is one of the most challenging and fundamental problem in astronomy. From the late eighteenth century, when numerous galaxies were being discovered as the telescopes became more and more powerful; cataloging of the celestial objects began[1]. The situation became much more complicated after the discovery of quasars. A quasar is a quasi-stellar radio source which emits mainly radio waves and visible lights, but to some extent ultraviolet rays, infrared waves, X-ray and gamma rays. Identification of quasars are usually done by identifying characteristic high ionization emission lines in optical spectrum along with spectroscopic follow up of optical sources having a radio counterpart[2–4] . They are very difficult to distinguish from the stars based on telescopic observation with manual template matching as the quasar are at least few billion light years away from the earth. In addition, the data size of source catalogues increases exponentially as the new facilities like the square kilometer array (SKA) and the Large Synoptic Array Telescope (LSST) are capable to catalogue billions of stars and galaxies. Hence, automated techniques are needed to classify the celestial objects.

While there can be large number of algorithms capable of automating the job of classification of celestial objects a selected few have been chosen for the exercise. These choices are based on mainly two aspects of the model: a) simplicity, b) performance. A multiclass logistic model is the most straight forward model where a logit consisting of a linear combinations of independent variables is passed to a logistic function to make the prediction about the target variable. In search of better performance more complex and inherently nonlinear models have been explored. Tree based models make successive splits based on threshold values of independent variables to maximize information gain for the splits. In this work decision tree[5] and it's extremely boosted version, namely xgboost[6] , have been tested. On the other hand a neural network[7] works by activating various layers of neurons based on a predetermined class of activation function for every layers of neurons in the model. A nearest neighbor method of classification works[8] by embedding the data in higher dimension and finding Minkowski distance from the instance point to the neighboring points of various class of data. In search for performance a naive Bayes model[9] was also considered which classifies using statistical inference from independent variables.

Naive Bayes, logistic regression and nearest neighbor methods are somewhat simplistic and can be explained easily but that simplicity comes with a tread off with performance. To compensate that complex and nonlinear models like decision tree, XGBoost and neural networks have been also taken under examination.

2. DESCRIPTION OF DATA

The data set contains 10,000 instances where each instance is a collection of 17 fields containing information about a particular celestial object and a final information about the class of the object under consideration. Let us look into these 17 pieces of information.

The first column is objid, this is an object identifier. The second and the third columns are right ascension and declination respectively. These two columns are representative of longitude and latitude in the sky. While pin-pointing a celestial object in the night sky these are very important. The next five columns are named as u,g,r,i and z respectively. Here the letters stand for:
u: ultraviolet2
g: green
r: red
i: infrared
z: near-infrared.
As the names suggest these are the names of the filters used. The numbers in the respective columns represent the normalized intensity of the radiation coming from the celestial object in the corresponding color bands.

Next four columns are run, rerun, camcol and field. These are the scan number, re-scan number, camera and field information.

Redshift column contains data about the relative velocity of the object compared to us. If an object moves away from us the frequency of its emitted light gets reduced and if it moves closer to us it's frequency is increased. As the current widely accepted model of universe depicts that all celestial objects are moving away from each other due to a phenomena called Hubble expansion redshift or reduction of frequency is prominent in almost all object that has their own light. The plate number is the identifier of the circular plate that was held at the focal point of the telescope. "mjd" can be thought of date in some form where as 'fiber id' is simply the identification of the optical fiber connected to the plate to record the data.
Since the information about scan number, camera, field or fiber id are irrelevant to differentiate the celestial objects, we can safely drop the column describing objid, run, rerun, camcol, field, plate, mjd and fiber id.

3. DESCRIPTION OF DATA

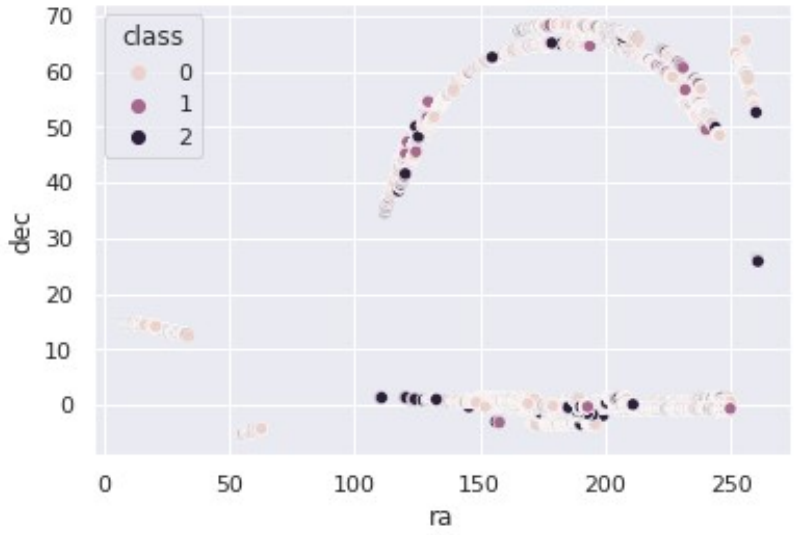

Figure 1:
Distribution of celestial bodies with respect to ascension and declination

Preprocessing of this data is kind of trivial as the data has no null values. Inspecting unique values in the columns show that “objid” and “rerun” column has only one value so we dropped it as the information becomes useless. The ‘specobj’ column has very high values so we normalized it own. All the features are numerical except the target variable hence the target variable was label encoded. Then all the independent variables were standardized using standard scalar. We refrained from removing outliers as that would remove some of the useful instances based on the fact that quasars have high “redshift” but others have very less value for it. While finding for any patterns we first dived into whether the position of the object with respect to our earth has anything to do with it. So we plotted its declination vs right ascension and found that there was no pattern in the appearance of class in the data [Fig:1].

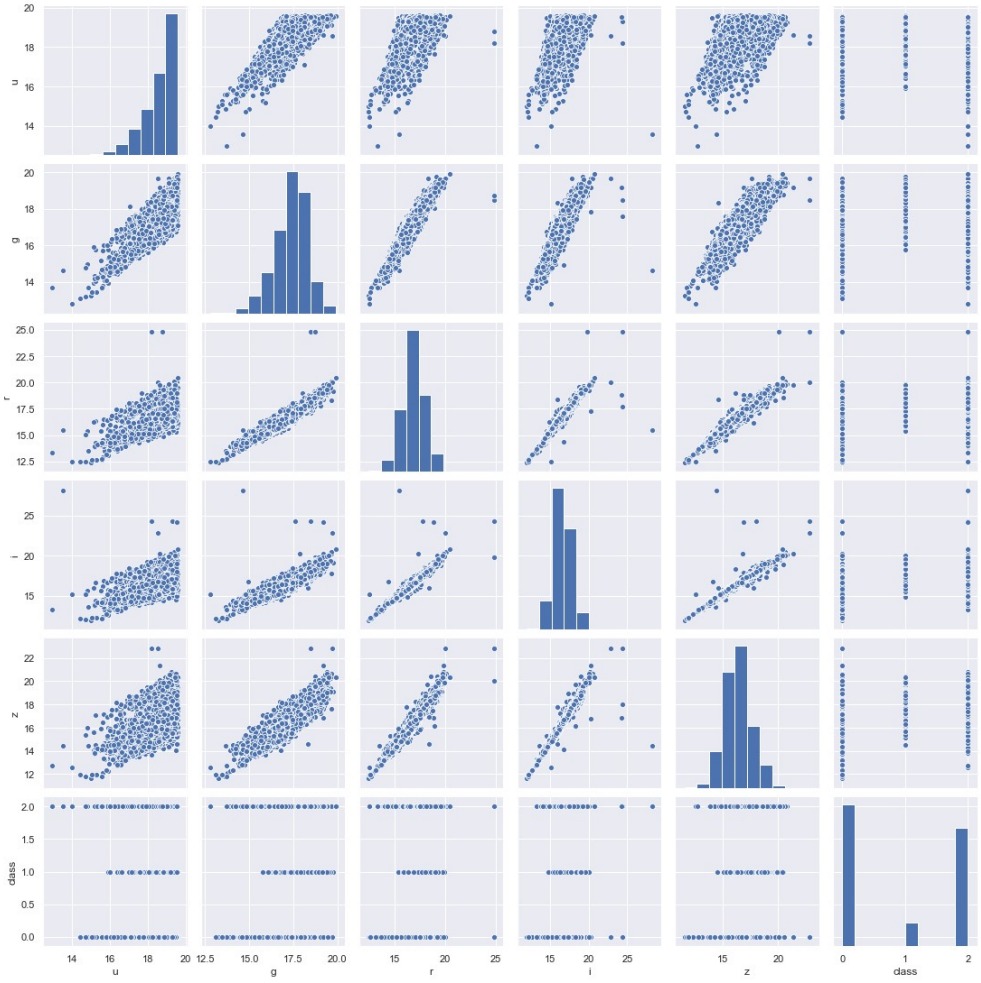

Figure 2: pair plots of radiation profiles

It is evident from the figure that angular position of earth is not at all related to star classification. The next five columns are very very important for the classification as ‘u’, ’g’, ’r’, ’i’, ’z’ gives the shape of the radiation profile of the data; one can think of it as a profile of black body radiation coming from the object. These profiles are often dictated by the process of energy generation which, in turn, dictate the classification of the object. To visualize the correlation within the five columns, pair plots [Fig:2] are done which shows some of the pairs are highly correlated.

The following plot [Fig:3A] describes the magnitudes of radiation in the two extreme filters namely ultraviolet and near-infrared of the three different classes. Here we can see that similar celestial objects appear in increasing trends but they form different parallel streaks. Class 0 is the lower most to radiate in near-infrared band, class 1 and 2 are separate but class 2 appears in two different linear profiles where as class 1 is sandwiched between those the class 2 profiles. As

these five profiles are very closely related to each other by Plank's distribution function any other combination of these columns results into a very similar situation.

Figure 3A: Plot describing the relation between magnitudes of radiation in the two extreme filters namely ultraviolet(u) and near-infrared(z)

If we move forward along the features we can comprehend that 'run', 'rerun', 'camcol' and 'field' are related to how the initial measurement was done and the inherent classification of the celestial object is not at all related to these features. Similarly by the same arguments, the features, named 'plate', 'mjd' and 'fiberid' can be discarded. But the remaining feature called 'redshift' is very important. If we take any of the frequency band response and redshift and plot them together we can see that objects of class 1 has much more higher velocity compared to the other twos [Fig:3B] .

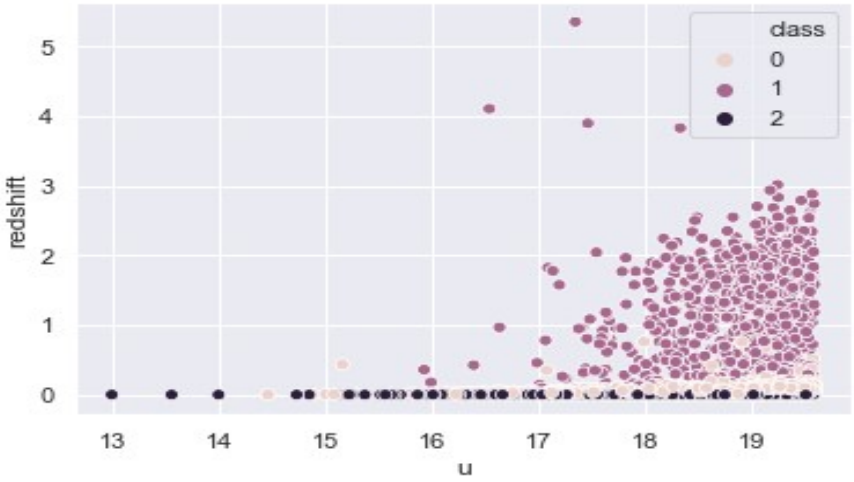

Figure 3B: Plot describing redshift as a function of intensity received through ultraviolet channel.

We can incorporate all of these knowledge in a single display if we plot the data in a 3D manner [Fig:4]. The green separate blob of data denotes the class 1 object and class 0 and class 2 form two

parallel streaks in the ‘z-u’ plane. As we know that quasars are high velocity objects compared to stars and galaxies our model can leverage that information.

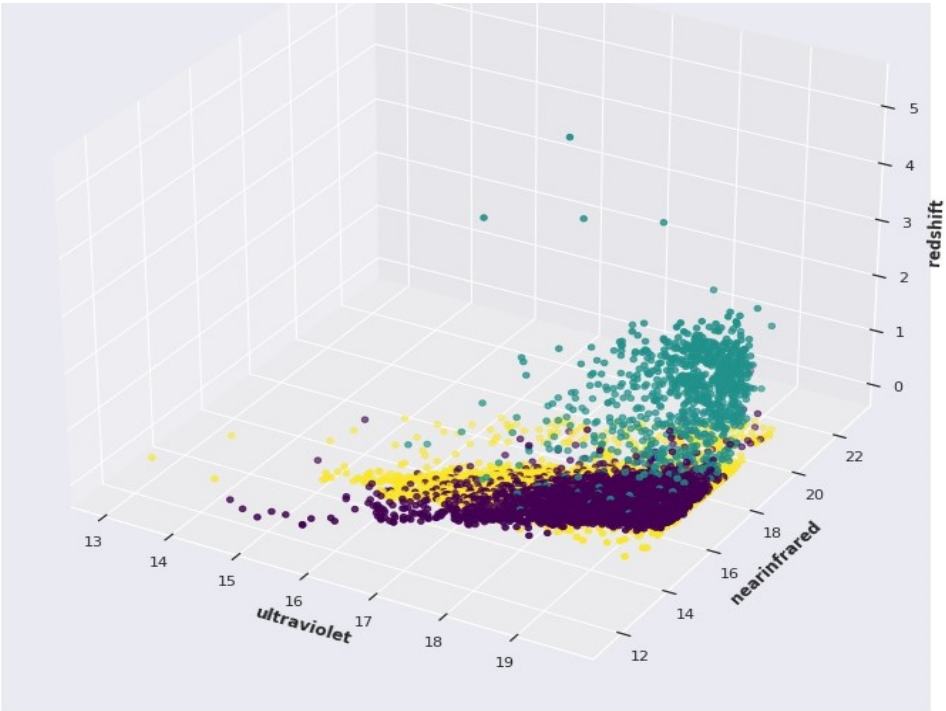

Figure 4: ultraviolet-nearinfrared-redshit plot of 3 classes

4. MACHINE LEARNING MODELS

Based on various attempts to model the class of an illuminating celestial object we found that feeding unnecessary data to the model actually hampers the prediction accuracy, so all the features that is related to how the measurements were done is dropped and only those features (namely ’u’,’g’,’i’,’r’,’z’ and ’class’) were kept which represent any property of the object itself.

After embedding the 6 dimensional data into various dimensions using Principle Component Analysis (PCA) and considering the confusion matrices it was seen that maximum accuracy is achieved when n-components is equal to 6. If a 6 dimensional data gives best result for n-components=6 in PCA then it is meaningless to use PCA. Hence the preprocessing was done by just using standard scalar on the independent features.

The data was divided randomly into 5 parts, 4 of which was used to train the data and the final one was reserved to test the model. Instead of using a fixed test data set, cross validation scheme has been used to rule out over fit or under fit the training data. To assess the performance of any machine learning model, confusion matrix is calculated which gives the value of precision, recall and F1 score. Precision value indicates how good the model to identify true positive whether recall value indicates how good the model to discard false negative. F1 score is the harmonic mean of precision and recall value which signifies overall performance of the model.

Here we have considered various algorithms. To begin with we used logistic regression which provided around 95 percent accuracy. As all the variables are continuous in nature, maximum number of iterations to find convergence of the model has to be increased. The accuracy, confusion matrix and classification report turned out to be acceptable. In search for whether its accuracy can be increased any further we considered a few more algorithms: decision tree, naive Bayes, nearest neighbor, simple neural network and XGBoost algorithm. The decision tree uses gini index to measure information gain. The naive Bayes algorithm uses Gaussian function as kernel and Bayes’ theorem to find the class. The nearest neighbor algorithm we considered here, uses generalized Minkowski distance to measure whether a data point is neighbor to another or not. One can simply conform to Euclidean distance as well but generalization

provides a scope of improvement. In the neural network we took 6*6=36 neurons in the first hidden layer and 6 neurons in the second hidden layer; as the input data have 6 independent variables 36 neurons were taken in the first layer to allow all possible kinds of second order interactions. The activation function of the network was taken to be ReLU. Apart from accuracy of the model, F1 score of the models are also checked. Comparisons among the models on the basis of precision, recall, F1 score and cross validation score are as following:

Table 1. Comparison among the classifiers on the basis of precision, recall and F1score

Classifier	Class	Precision	Recall	F1 score	Cross validation score
Logistic Regression	0	0.94	0.98	0.96	[0.9575, 0.9615, 0.9615, 0.9535, 0.957]
	1	0.92	0.96	0.94	
	2	0.99	0.93	0.96	
Decision tree	0	0.98	0.99	0.99	[0.9895 0.9855 0.9875 0.984 0.988]
	1	0.96	0.93	0.93	
	2	1.00	1.00	1.00	
Naive Bayes	0	0.98	0.99	0.98	[0.983 0.9845 0.9835 0.9735 0.979]
	1	0.94	0.91	0.93	
	2	0.99	0.99	0.99	
Nearest Neighbor	0	0.93	0.97	0.95	[0.9835 0.985 0.985 0.975 0.979]
	1	0.93	0.96	0.95	
	2	0.98	0.92	0.95	
Neural Network	0	0.97	0.99	0.98	[0.98 0.985 0.985 0.978 0.9855]
	1	0.92	0.96	0.94	
	2	1.00	0.98	0.99	
XGBoost	0	0.99	0.99	0.99	[0.993 , 0.993 , 0.9915, 0.9875, 0.9905]
	1	0.94	0.96	0.95	
	2	1.00	1.00	1.00	

It is evident from the above table that most of the classifiers work efficiently to identify class 2 object but the efficiency drops to segregate class 0 and class 1 objects. The reason can be understood if we look [Fig:3b] and [Fig:4]. Decision tree and XGBoost classifiers, both of them show 99% and 100% accuracy to identify class 0 and class 2 object respectively, but XGBoost classifier performs slightly better and more consistent than decision tree to identify class 1.

5. DISCUSSION

From the results in the previous sections it has been seen that machine learning models can provide an automated methodology to process huge amount of telescopic data about various kinds of celestial objects and their classification. While classifying an celestial object, even the most rudimentary model: multiclass logistic regression can give an F1 score of 0.95. The level of performance can then be further enhanced by deploying more sophisticated and complex models. Among these models a decision tree gives a combined F1 score of 0.96. When numerous trees are bagged together to make a random forest model the score reaches 0.98. A boosting method using a decision tree as a base model beats the random forest by providing an F1 of 0.99. After cross validating all the models on the whole set of data it has been observed that XGBoost model consistently out performs all the other models. This model can be deployed in real time data accusation to detect the celestial objects on the go.

REFERENCES

[1] *William Herschel, "Catalogue of a Second Thousand of New Nebulae and Clusters of Stars; With a Few Introductory Remarks on the Construction of the Heavens", Philosophical Transactions of the Royal Society of London , vol. 79,* ***(1789)****, pp. 212– 255 .*

[2] *M. Schmidt, " A Star-Like Object with Large Red-Shift ",* Nature, vol. 197, (**1963**), pp. 1040.

[3] *D. W. Weedman, " Seyfert galaxies", Annual Review of Astronomy and Astrophysics, vol. 15, (**1977**), pp. 69-95.*

[4] *R. Antonucci, " A panchromatic review of thermal and nonthermal active galactic nuclei ", Astronomical & Astrophysical Transactions, vol. 27, (**2012**), pp. 557-602.*

[5] *T. Spinka, T. Carpenter, R. J. Brunner, R. Aydt, L. Auvil, T. Redman, D, Tcheng, " Quasar Identification and Classification with Decision Trees ", American Astronomical Society Meeting Abstracts, vol 203, (**2003**), id.04.09;Bulletin of the American Astronomical Society, vol. 35, (**2003**), pp.1208.*

[6] Tianki Chen, Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (**2016**), pp. 785-794.

[7] Edward J. Kim, Robert J. Brunner, "Star–galaxy classification using deep convolutional neural networks", *Monthly Notices of the Royal Astronomical Society*, vol. 464, (**2017**), pp. 4463–4475.

[8] LiLi Li, YanXia Zhang & YongHeng Zhao, "*k*-Nearest Neighbors for automated classification of celestial objects", Sci. China Ser. G-Phys. Mech. Astron., vol. 51,(**2008**), pp. 916-922.

[9] Marc Henrion, Daniel J. Mortlock, David J. Hand, Axel Gandy,"A Bayesian approach to star–galaxy classification", *Monthly Notices of the Royal Astronomical Society, vol. 412, (**2011**), pp. 2286–2302.*

4.

Author Name(s) and Affiliation(s)

Author names and affiliations are to be centered beneath the title and printed in Times New Roman 12-point, non-boldface type. (See example below)

4.1. Affiliations

Affiliations are centered, italicized, not bold. Include e-mail addresses if possible.

For example:

Author[1], Author[2] and Author[3]

[1]Affiliation
[2]Affiliation
[3]Affiliation
[1]Email, [2]Email, [3]Email

7.1. Tables

Place tables as close as possible to the text they refer to and aligned center. A table is labeled *Table* and given a number (*e.g.*, **Table 1. Sample Datasheet with Attributes in Linguistic Term**) it should be numbered consecutively. The table label and caption or title appears 9pt space above the table, 6pt space after the text or paragraph if any; it should be uniforms fonts and font size, and use 11pt font size and Helvetica style, capitalized similar to paper title, aligned center and bold face. Sources and notes appear below the table, aligned left. All tables must be in portrait orientation.

For Example:

Table 1. Table Label

7.2. Figures

Place figures as close as possible to the text they refer to and aligned center. Photos, graphs, charts or diagram should be labeled *Figure* (do not abbreviate) and appear 6pt space below the figure, 12pt space before the next text or paragraph, and assigned a number consecutively. The label and title should be in line with the figure number (*e.g.*, **Figure 1. Location Error Rate of Three Schemes**), it should be uniforms fonts and font size; use 11pt font size and Helvetica style, capitalized similar to paper title, aligned center and bold face. Source (if any) appear underneath, flush left. Figures should be at good enough quality. Minimum image dimensions are 6 cm (2.3622 in) wide by 6 cm (2.3622 in) high.

For Example:

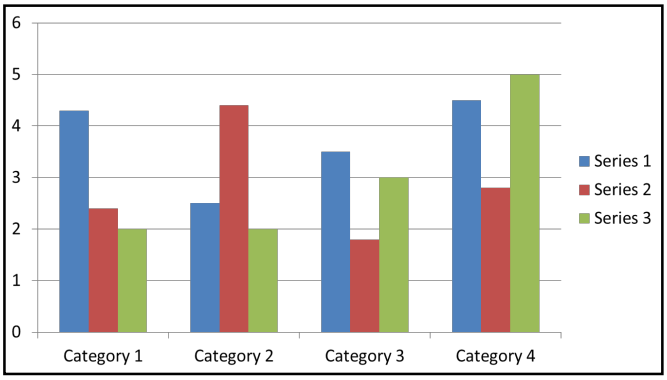

Figure 1. Figure Label

7.3. Equations

Including symbols and equations in the text, the variable name and style must be consistent with those in the equations. Equations should be indented at the left margin and numbered at the right margin, equation number is enclosed with open and close parenthesis () Time New Roman in style and 11pt font size. Define all symbols the first time they are used. All equation symbols must be defined in a clear and understandable way.

For Example:

$$\varphi_{\mu\nu}(z) = \frac{\left\|k_{\mu\nu}\right\|^2}{\sigma^2} e^{-\frac{\left\|k_{\mu\nu}\right\|^2 \|z\|^2}{\sigma^2}} [e^{ik_{\mu\nu}z} - e^{\frac{-\sigma^2}{2}}] \quad (1)$$

8. First-order Headings

For example, "**1. Introduction**", should be Times New Roman 13-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

8.1. Second-order Headings (Sub-heading)

As in this heading, they should be Times New Roman 11-point boldface, initially capitalized, flush left, with one blank line before, and one after.

8.1.1. Third-order Headings: Third-order headings, as in this paragraph, are discouraged. However, if you must use them, use 11-point Times New Roman, boldface, initially capitalized, flush left, and proceeded by one blank line, followed by a colon and your text on the same line.

9. Footnotes

Use footnotes sparingly (or not at all) and place them at the bottom of the column of the page on which they are referenced to. Use Times New Roman 9-point type, single-spaced. To help your readers, avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence).

Appendix

An appendix, if needed, should appear before the acknowledgments.

Acknowledgments

These should be brief and placed at the end of the text before the references.

REFERENCES

List and number all bibliographical references that has important contribution on the paper, (if possible, limit to 30, which only are necessary citations are recommended). 9-point Times New Roman, fully justified, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [1]. Do not abbreviate the months. Don't forget to put period (.) at the end of each reference. (See examples below)

11.1. Journal Article

[10] *C. D. Scott and R. E. Smalley, "Diagnostic Ultrasound: Principles and Instruments", Journal of Nanosci. Nanotechnology., vol. 3, no. 2,* ***(2003)****, pp. 75-80.*

11.2. Book

[11] *H. S. Nalwa, Editor, "Magnetic Nanostructures", American Scientific Publishers, Los Angeles,* ***(2003)****.*

11.3. Chapter in a Book

[12] *H. V. Jansen, N. R. Tas and J. W. Berenschot, "Encyclopedia of Nanoscience and Nanotechnology", Edited H. S. Nalwa, American Scientific Publishers, Los Angeles, vol. 5,* ***(2004)****, pp. 163-275.*

11.4. Conference Proceedings

[13] *J. Kimura and H. Shibasaki, "Recent Advances in Clinical Neurophysiology", Proceedings of the 10th International Congress of EMG and Clinical Neurophysiology, Kyoto, Japan,* ***(1995)*** *October 15-19.*

11.5. Patent

[14] *C. E. Larsen, R. Trip and C. R. Johnson, "Methods for procedures related to the electrophysiology of the heart", U.S. Patent 5,529,067,* ***(1995)*** *June 25.*