As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 9(05): 35-41, 2023
Copyright © 2023International Journal for Modern Trends in Science and Technology
ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0905006
Available online at: http://www.ijmtst.com/vol9issue05.html

An Analysis of Machine Learning Methods for the detection ofExoplanets

Karan Chawla

Ashoka Unviersity

To Cite this Article
Karan Chawla. An Analysis of Machine Learning Methods for the detection ofExoplanets. International Journal for Modern Trends in Science and Technology 2023, 9(05), pp. 35-41. https://doi.org/10.46501/IJMTST0905006

Article Info
Received: 28 March 2023; Accepted: 02 May 2023; Published: 05 May 2023.

ABSTRACT

Over a million stars were observed over the past ten years in an effort to findtransiting planets. Manual interpretation of prospective exoplanet candidatesrequires a lot of work and is prone to human mistake, with difficult to measureoutcomes. Thanks to extensive ground- and space-based searches, the number of newly discovered planets and planetary systems has increased from a single planetto hundreds. The discovery of individual planets is no longer the main focus ofscience; instead, it is characterizing planets, which entails searching for thepresence of various chemical elements in the tantalizing hope of finding planetswith habitable environments and locating biological life. The perfectplanet-detection algorithm should be quick, noise-resistant, and capable of learningand abstracting extremely nonlinear systems. The best platform is a neural networktrained with simulated data to identify planets. A computational method calleddeep learning with a neural network attempt to simulate the biological processthrough which a brain solves issues by connecting groups of neural units. Mostinformation about the planet is unlabeled, including information obtained viatransit spectroscopy. Automating the categorization of habitability can help with this problem's solution. SMOTE, or Synthetic Minority Oversampling Technique,was the approach that best addressed the problem of class imbalance. This methodproduced good results when compared to Random Oversampling and RandomUndersampling. Clustering and Anomaly Detection can be useful preparatoryprocedures for a complete inversion analysis utilizing the k-means clusteringapproach because none of them need understanding of the fundamental physics andchemistry of the observed atmospheres. This review paper assesses three machineslearning methods namely, Convolutional Neural Network, K-means clustering andSaha Bora activation function for the detection of exoplanets.

KEYWORDS:Artificial Intelligence, Deep Learning, K-means clustering, Saha-Bora Activation Function, Clustering.

1. INTRODUCTION

The discovery of planets outside of our solar system has advanced from a single planet to hundreds of newly discovered planets and planetary systems thanks to extensive ground- and space-based searches. Today, the focus of scientific inquiry has shifted from the discovery of individual planets to the characterisation of planets, which involves looking for the existence of various chemical components in the tantalizing hope of discovering planets with livable settings and identifying signs of biological life. When data is convolved using the best filter, the SNR of a transit detection may be maximized. Kernels are manually created to roughly represent what a human user would consider to be the best filter because it is not possible to analytically solve for the ideal filter in the situation of variable transit shapes.

Convolutional neural networks (CNN) have been utilized in the past to overcome comparable kernel optimization issues using deep learning. The thresholds that now constrain photometric surveys will be raised by future planet-hunting surveys like TESS, PLATO, and LSST by sampling brighter stars at quicker cadences and over bigger fields of views. Most methods to detect planets use a least squares optimization, grid-search, or matching filter strategy to increase the correlation between data and a straightforward transit model.

Using least-squares optimisation, the mean-squared error (MSE) between data and a model is sought to be as small as possible. Practical applications rely on a variety of numerical inversion approaches, which lately have been using more and more cutting-edge statistical and machine-learning (ML) techniques to increase the precision, accuracy, and speed of the executed retrievals. The last point is more important in light of the enormous number of exoplanet transits that are anticipated during the future years. The perfect planet-detection algorithm should be quick, noise-resistant, and capable of learning and abstracting extremely non-linear systems.

The best platform is a neural network trained with simulated data to identify planets. A neural network used in deep learning is a computer method for simulating the biological process through which a brain solves issues by connecting groups of neural units. Layers of "neurons" make up deep nets, and each layer is assigned a distinct weight to represent the relative weights of the various input parameters.

2. CONVOLUTIONAL NEURAL NETWORK

Decision Trees, Support Vector Machines, Logistic Regression, Random Forest Classifier and the Convolutional Neural Network are the The baseline algorithms for building upon the ensemble-CNN algorithm. Decision trees are graphs that resemble trees and include nodes that represent the points where one picks an attribute and poses a question. While leaves indicate the actual output or class label, edges are understood to represent the responses to the inquiry. With straightforward linear decision surfaces, decision trees are used in nonlinear decision-making. Creating training models that might be helpful in predicting the class or value of the target variable is essentially what decision trees are concerned with. Simple decision rules that have been learned from training data are the basis of this supervised learning system. It is taken into account starting at the tree's base when predicting a class label for a record. The root attribute and the record attribute values are contrasted. The branch corresponding to the value is followed, and the next node is taken into consideration, based on the comparison. The root node of the decision tree is divided into decision nodes. A decision node might further divide into terminal nodes (leaf nodes) or other decision nodes. You may consider each node in the tree as a test case for one or more attributes. Similar to this, each edge that descends from the node may be viewed as a potential solution to the test case. Every subtree anchored at the new node goes through this recursive procedure once again. Support vector machines are supervised learning-based machine learning techniques. They offer a practical method for classifying and doing regression analysis on data. SVM was created in the 1990s with the intention of offering non-linear approaches to data analysis. One of the most used machine learning algorithms is SVM. This is so that SVM can perform better than other model strategies even with a small number of features. When compared to the model's error, the SVM model is rather resilient. Additionally, SVM takes less time to compute than other models like neural networks. Finally, SVM is more efficient than the majority of models. The Input Layer, the Hidden Layer, and the Output Layer are the three layers that make up the SVM architecture.

Layer 1, also known as the input layer, contains training examples that are coupled to the hidden layer for processing the learning prediction. The output layer is also linked to this layer. Given that the dependent variable is dichotomous, the machine learning method known as logistic regression is employed to do regression analysis. It is a type of predictive analysis that's used to characterize data and ascertain how dependent binary variables relate to independent variables. As a result, it may be stated that the dependent variable is a binary variable, denoted by the data 1 (for yes, happy, success, etc.) or 0 (for no, sad, failure, etc.). P(Y = 1) is predicted by the logistic regression model as a function of X. Since probabilities may be predicted directly, logistic regression has an advantage over linear regression. Additionally, it keeps the training data's marginal probabilities. With logistic regression, there are several presumptions. Variables that need huge sample numbers or don't have any significance are not taken into account. It is mostly used to forecast binary output. The Random Forest classifier is essentially an algorithm for learning from ensembles of trees. As a result, it is made up of several separate decision trees that work together as an ensemble. In order to identify the final class of the test item, the classifier aggregates the votes from several decision trees from a portion of the training set that was randomly chosen. Every tree has a class prediction, therefore the prediction made by the model belongs to the class that receives the most votes. This also shows that if a large number of reasonably uncorrelated models work together as a committee, they will perform better than their individual component models. A single perceptron is multiplied by a weight and added to in the Perceptron Algorithm. However, there are several linear layers in Multilayer Perceptron (MLP). An input layer, a hidden layer, and an output layer are all components of a three-layer network. The input layer receives data, while the output layer produces results. Depending on the issue, the number of concealed layers can be changed. Every perceptron in MLP is connected to every other perceptron, which is referred to as being completely connected. Fully linked MLP has an excessive number of parameters, which might make weights difficult to control. Overfitting, inefficiency, and redundancy may come from this.

Convolutional Neural Networks (CNN) can handle more parameters and are more reliable than MLP. Convolution operations are carried out by filters in a convolutional neural network, which is a type of deep neural network. The foundational component of a CNN is the convolutional layer. The parameters of the layer are made up of several learnable filters that combine a modest responsive field while extending over the whole depth of the input volume. The forward pass processes the dot product between the filter entries and the input to create a 2-dimensional activation map by convolving each filter over the width and height of the input volume. As a result, the network develops filters that activate when it detects a specific type of feature at a specific geographical location in the data. The information and its result are convolved to the next layer via convolutional layers. This closely resembles how a cell in the visual brain might respond to a specific stimulus. Information specific to each convolutional neural procedure's receptive field. A CNN is made up of many kinds of layers. In order to predict the class probabilities for each feature, the convolutional layer builds a feature map. Applying a filter that scans the entire image does this. The amount of data that the convolutional layer generated for each feature is scaled down and the most important information is maintained by a pooling layer. The completely linked input layer is responsible for transforming the outputs from previous layers into a single vector that can be used as input for the layer after it. The feature analysis-generated input is subjected to weighting by the fully linked layer. This function's goal is to correctly forecast the label. The final probabilities for selecting an image class are produced by a fully linked output layer. A method of integrating various machine learning models called stacking. The term "meta learner" is first used in stacking as an alternative to the term "voting," which is frequently used in bagging. The primary function of stacking is to recognise or categorize the trustworthy model and meta learner, and it aids in the search for a technique for fusing the output of the best base-learner. The predictions of the metamodel input and base model, respectively, are Level-1 models and Level-0 models. In essence, stacked learners are employed for classification, where each occurrence inputs a class value prediction into a Level-0 model. These forecasts add to level 1 and come together to form the final forecast. Although it seems challenging, training the stacking model is not as complex as it seems. The stacking model's training procedures are quite similar to those of k-fold validation. The dataset is divided into two sets for this ensemble technique: the Train set and the Test set. But throughout the training phase, the test set is not used. The training set has k-number of folds added to it. If the input dataset has N data points, then these folds contain N/k number of points. It predicts the value of fold using the M number of models, and the M-Number of predictions is derived from the N/k data points. These forecasts may be sent into the meta learner, and the metal learner can forecast the outcomes.

3. K-MEANS CLUSTERING

Both the forward problem and the inverse problem fall under the umbrella of supervised learning (multivariate regression) in machine learning, where one attempts to predict a collection of objectives (outputs) given a set of characteristics (inputs). The selection of the variables to be considered as features and goals accounts for the majority of the differences between and. Any such supervised learning assignment must have access to high-quality labeled training data in order to be successful, which in turn necessitates an accurate simulation of the intricate radiative transfer process. In other words, the forward model's quality and accuracy will determine how well the inversion performs, and as a result, the inversion outcomes are model-dependent. In contrast to the majority of the transit spectroscopy literature that makes use of machine learning techniques, which focuses on supervised learning, in this paper one can approach the inversion problem from the point of view of unsupervised learning, i.e The focus will remain exclusively on the set of measured wavelength-dependent modulations, without any reference to the corresponding atmospheric parameters. In other words, the paper's whole numerical work will be carried out in an entirely unsupervised manner. Although it is not necessary to have prior knowledge of the labels (the atmospheric parameters), doing so may make it simpler to see and comprehend the results and to relate them to the underlying physics and chemistry.

Unsupervised machine learning techniques make an effort to delve deeper into the underlying structure of the data set, revealing hidden patterns, correlations, and linkages. This approach offers a more exploratory, open-minded view of the data in anticipation of (and as a prelude to) any supervised learning tasks that may come next. Data wrangling, preprocessing, initial exploratory data analysis based on summary statistics, factorization techniques like principal component analysis, dimensionality reduction, and manifold learning, and grouping techniques, which try to identify similar groups or anomalies, are the only unsupervised learning tasks that are most pertinent for planetary spectroscopy that will be covered in this paper. Another common unsupervised job is clustering. By using the K-means clustering technique to divide a library of synthetic spectra into distinct groups, researchers have employed clustering in the context of planetary transmission spectroscopy to derive informed priors for the radiative transfer retrieval model. The primary goal was to increase retrieval speed by refining the first parameter estimation. The number of classes to be employed in the technique is chosen based on this rationale.

While in principle, the more classes, the better, the best number of classes relies on the data's degree of noise, which may result inmisclassification, in practise, the best number of classes is dependent on the noise level. The examination of the data identified a number of unique groupings (branches) that correspond to various classifications of atmospheres. In order to locate the individual branches, which are the important regions in terms of physics and chemistry, rather than to discover excellent beginning hypotheses, this means that our purpose is using the clustering approach. From our perspective, the number of classes is determined by the number of intriguing physics and/or chemistry regimes, which in our instance is correlated with the number of gas components present in the atmosphere, rather than the resolution or noise level.

The scikit-learn K-means clustering tool may be used to examine the benchmark data set. The clustering is carried out in the entire 13-dimensional PCA space following the standardization of the PCA component parts. Without any prior information of the temperature and makeup of the atmosphere, the classification was performed using the K-means method, which accurately distinguishes between the branches for clouds, water, HCN, and ammonia. Future studies in this area have a potential path forward because of the effectiveness of the unsupervised clustering method.

There are several clustering algorithms that can determine the ideal number of clusters on their own, unlike the K-means method, which needs the number of clusters to be predetermined. As an alternative, one might examine the ideal number of clusters within K-means itself, using a method like silhouette analysis, for example. The key takeaway from this experiment is that the spectrum data clustering contains significant information regarding the presence or absence of certain atmospheric elements. One of these physics-driven clusters may be promptly linked to a newly detected spectrum, allowing the planet to be quickly categorized based on the makeup of its atmosphere.

4. SAHA-BORA ACITVATION FUNCTION

Artificial neural networks, or neural networks, are a set of linked units arranged in layers that process information signals by dynamically reacting to inputs.

The network's layers are arranged so that inputs go to the input layer and output comes from neurons in one or more hidden layers after processing. Computing neurons make up the hidden layers, which are linked to the input and output layers via a network of weighted connections. With each input provided to the network, weights are changed so that the error between the desired and observed output is kept to a low. This is how the network learns from input patterns. Hidden layers have a unique capability called activation function that allows neurons to process signals and spread them across the network. The computation of the discrepancy between the observed and intended output is the responsibility of a particular kind of ANN termed back propagation, which then feeds this discrepancy back to the network with each cycle or 'epoch'. The weights are adjusted appropriately, and the network is trained or learned until the error is minimized. A functional mapping between inputs and outputs is provided by the activation function. The network can learn from and model complex datasets including audio, video, and text because of this. The Sigmoid, hyperbolic tangent, and Relu activation functions are the most common. Significant accomplishments are suggested by the analysis on quasar-star classification using machine learning and the design of unique activation function for exoplanet classification. They suggest an activation function that needs less effort to tune the hyperparameters than the conventional activation function. It has also been demonstrated to be a successful answer to the first-order differential equation. In an effort to get successful results, several have experimented using ANN in habitability categorization challenges. The categorization of exoplanets has since been developed using a unique elastic KNN model. The authors claim that this model can handle a wide range of input parameters while maintaining global optimum. The new activation function features an optima and will be utilized to train a neural network for habitability categorization. Evidently, compared to the more popular sigmoid function, there is less flattening of the function in the graphic simulations shown below. As a result, the formulation should make it easier to deal with local oscillations. A first order polynomial can be used to approximate the variable term in the SBAF denominator. This could enable us to avoid pricey floating-point calculations without sacrificing precision. The maxima's exclusivity within the specified timeframe must be demonstrated. This will get around the local maximum issue.

5. ANALYSIS

For the convolutional neural network, automating the categorization of habitability, machine learning can help with the issue of habitability disposition. Synthetic Minority Oversampling Technique (SMOTE) was the method that dealt with the issue of class imbalance the best. In comparison to Random Oversampling and Random Undersampling, this approach yielded good results. After boosting the minority class data by synthetically creating new examples, the machine learning model was able to train and generalize successfully. Support Vector Classifier was also used to achieve cost-sensitive learning, but the outcomes were the same as those attained by employing SMOTE. The limitation of the study is that it is not easy to detect reflected light from a planet’s atmosphere. For the k-means clustering method, Exhaustive tasks like scanning transiting light curves for planetary signals are jobs that ML techniques are capable of handling. One of the most challenging characteristics that prevent ML methods from performing to their full potential is noise in the light curve signals. By producing false positives or even obscuring the transit signals from the detection models, noisy features might trick AI algorithms. Human interaction is still necessary (for example, in feature extraction) even if the current ML algorithms lighten the workload for scientists working to validate exoplanet discoveries. Additionally, weak transit signals offer a fantastic chance to discover exoplanets that resemble Earth. The optimal machine learning model should be able to analyze weak signals. To solve the issue posed by transits seen in low SNR light curves, this calls for a better grade of detection and identification capability. For these reasons, MRA appears to be a viable method for finding tiny planets and validating the signals that are found. MRA may reduce the amount of the data while simultaneously collecting fine details from the light curves. This enhances the ML models' identification performance and considerably reduces the execution time.

6. CONCLUSION

With a 99.62% accuracy rate, the Ensemble-CNN approach exceeds the Transit approach, Radial Velocity, Direct Imaging, and Gravitational Microlensing. Because none of them necessitate knowledge of the underlying physics and chemistry of the observed atmospheres, Clustering and Anomaly Detection can serve as helpful preprocessing steps for a thorough inversion analysis using the k-means clustering method. Strong correlation between the spectral data has been demonstrated, calling for the usage of low-dimensional representations. Therefore, dimensionality reduction techniques have been applied. Intriguing data structures have also been discovered via research that have distinct branch topologies and correlate to varied chemical regimes. The technique of classifying habitat appropriateness is challenging. Despite the abundance of sophisticated methods that mix supervised and unsupervised learning techniques in the literature, differentiating between the classes of psychroplanet and mesoplanet is incredibly difficult because of the fragile border between them. In a 2018 research, the performance of the Cobb-Douglas Habitability Score (CDHS) was compared to that of other machine learning algorithms, and its elasticity was studied. Given our limited knowledge of exoplanets and their habitability, these results and methodologies offer a key first step towards future ground- and space-based observatories automatically recognising objects of interest from huge databases. A framework for forward and backward pass training may be useful. The variable term in the SBAF is based on earlier modeling of topics like production functions and the usage of optimization theory in production economics. The development of classification algorithms may frequently be required to address data complexity or bias in order to improve the initial method, lessen class imbalance, or alter confidence intervals. This is the primary argument in favor of creating a special activation function for neural networks.

7. FUTURE DIRECTIONS

For the K-means clustering method, Future research should focus on assessing the performance of additional MRA methods, such as Ensemble Empirical Mode Decomposition, Stationary Wavelet Transform, Empirical Mode Decomposition (EMD), and employing the reconstructed signal in the exoplanet identification stage. Exoplanet transits differ in form due to factors like star activity, for example. Therefore, a straightforward template is insufficient to capture the finer details, especially when the signal is weaker than the noise or when there are significant systematics. To learn the photometric characteristics of a transiting exoplanet, we employ an artificial neural network. Millions of light curves may be processed in a couple of seconds using deep machine learning. In order to improve the detection resilience to noise, future research needs to use deep learning techniques like short term memory and PReLU. The network design needs to be optimized in order to be made to respond to particular challenges, which requires more study. By eliminating systematics from the time series, a pre-processing step might considerably enhance the performance of transit identification.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Maxwell, A.E., Warner, T.A., Fang, F.: Implementation of machine learningclassification in remote sensing: an applied review. Int. J. Remote Sens. 39(9),2784–2817 (2018)

[2] Singh, S.P., Misra, D.K.: Exoplanet hunting in deep space with machine learning.IJRESM 3(9), 187–192 (2020)

[3] Bentum, M.J.: Algorithms for direct radio detections of exoplanets in theneighbourhood of radiating host stars. In: IEEE Aerospace Conference, pp. 1–7(2018). https://doi.org/10.1109/AERO.2018.8396590

[4] Jagtap, R., Inamdar, U., Dere, S., Fatima, M., Shardoor, N.: Habitability ofexoplanets using deep learning. In: IEEE International IOT, Electronics andMechatronics Conference (2021)

[5] Malik, A., Moster, B. P., &Obermeier, C. (2020). Exoplanet Detection using Machine Learning. arXiv preprint arXiv:2011.14135.

[6] Koch, D. G., Borucki, W., Dunham, E., Geary, J., Gilliland, R., Jenkins, J., ... & Weiss, M. (2004, October). Overview and status of the Kepler Mission. In Optical, Infrared, and Millimeter Space Telescopes (Vol. 5487, pp. 1491-1500). International Society for Optics and Photonics.

[7] Goldilocks Zone. (2021, March 4). Exoplanet Exploration: Planets Beyond Our Solar System. https://exoplanets.nasa.gov/resources/323/goldilocks-zone /

[8] S. Dere, M. Fatima, R. Jagtap, U. Inamdar and N. B. Shardoor, "Anomaly Detection in Astronomical Objects of Galaxies Using Deep Learning", 2021 7th International

Conference on Advanced Computing and Communication Systems (ICACCS), 2021.

[9] R. Jagtap, U. Inamdar, S. Dere, M. Fatima and N. B. Shardoor, "Habitability of Exoplanets using Deep Learning", 2021 IEEE International IOT Electronics and Mechatronics Conference (IEMTRONICS), 2021.

[10] 0.Suryoday Basak, Surbhi Agrawal, SnehanshuSaha, Abhijit Jeremiel Theophilus, Kakoli Bora, Gouri Deshpande, et al., "Habitability Classification of Exoplanets: A Machine Learning Insight", 2018.

[11] BrychanManry, George Sturrock and Sohail Rafiqi, "Machine Learning Pipeline for Exoplanet Classification", 2019.

[12] Wahyono, M. A. Rahman and SN. Azhari, "Classification of Galaxy Morphological Image Based on Convolutional Neural Network", International Journal of Advanced Research in Science Engineering and Technology, vol. 5, no. 6, 2018.

[13] A. Mittal, A. Soorya, P. Nagrath and D.J. Hemant, "Data augmentation based morphological classification of galaxies using deep convolutional neural networks. (part of Springer Nature 2019)" in , Springer-Verlag GmbH Germany, 2019.

[14] A. Gauthier, A. Jain and E. Noordeh, "Galaxy Morphology Classification" in , Stanford University, 2016

[15] V. Lukic and M. Bruggen, "Galaxy Classifications with Deep Learning Hamburger Sternwarte International Astronomical Union" in , Hamburg, Germany:University of Hamburg, 2017.

[16] J. H. LL. Murrugarra and S. T. N. Hirata, "Galaxy image classification. Department of Computer Science Institute of Mathematics and Statistics" in , Sao Paulo, Brazil:University of Sao Paulo.

[17] Rajeev Mishra, "Predicting habitable exoplanets from NASA's Kepler mission data using Machine Learning", 2017.

[18] Megan Ansdell, YaniIoannou, Hugh P. Osborn and Michele Sasdelli, "Scientific Domain Knowledge Improves Exoplanet Transit Classification with Deep Learning", 2018

[19] Christopher J. Shallue and Andrew Vanderburg, "Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90", 2018

[20] Piyush Gawade, Akshay Mayekar, Ashish Bhosale and Sanjay Jadhav, "Finding New Earths Using Machine Learning & Committee Machine", 2020.

[21] S. Saha, S. Agrawal, R. Manikandan, K. Bora, S. Routh and A. Narasimhamurthy, "ASTROMLSKIT: A New Statistical Machine Learning Toolkit: A Platform for Data Analytics in Astronomy", April 2015.

[22] William J. Borucki et al., "Kepler planet-detection mission: introduction and first results", Science, vol. 327, no. 5968, pp. 977-980, 2010.

[23] A. Wolszczan, "Searches for planets around neutron stars", Celest. Mech. Dyn. Astr., vol. 68, pp. 13, 1997.

[24] J. Bloom and J. Richards, "Data mining and machine-learning in time-domain discovery & classification", Adv. Mach. Learn. Data Min. Astron., 2011.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886-893, 2005.