This document is currently being worked upon.

List of projects:

5. Stars-Galaxies and Stars-Quasars 7. Stocks 8. Proxima B 9. GALEX 10.Heat Equation for Stock Market Modeling

1 Exoplanets and Classification: Implicit ML approach

We consider the PHL-EC dataset. Led by the NASA Kepler Mission, around 3416 planets have been confirmed, and 4900+ celestial objects remain as candidates that are yet to be confirmed as planets. The discovery and characterization of exoplanets requires both, extremely accurate instrumentation and sophisticated statistical methods to extract the weak planetary signals from the dominant starlight or very large samples. Thus, two methods of synthetic oversampling are explored:

1. **By assuming a Poisson distribution in the data (surface temperature of exoplanets) and applying a novel hybrid algorithm to label synthetic data.** A novel method (hybrid SVM-KNN) to label synthetic data with 100% accuracy. This is required since the two classes, meso (26) and psychro planets (18) in PHL-EC have significantly lesser number of samples compared to the class non-habitable (3300 approximately)– **Artificially Augmenting Data in a Bounded Manner:** The challenge with artificially oversampling data in PHL-EC is that the original data available is too less to estimate a reliable probability distribution which is satisfactorily representative of the probability density of the naturally occurring data. For this, a *bounding mechanism* should be used so that while augmenting the data set artificially, the values of each feature or observable does not exceed the physical limits of the respective observable, and the physical limits are analyzed from the naturally occurring data.

   For this purpose, we use a hybrid of SVM and K-NN to set the limits for the observables. The steps in the SVM-KNN algorithm are summarized as: The best boundary between the psychroplanets and mesoplanets are found using SVM with a linear kernel. By analyzing the distribution of either class, data points are artificially created. Using the boundary determined in Step 1, an artificial data point is analyzed to determine if it satisfies the boundary conditions: if a data point generated for one class falls within the boundary of the respective class, the data point is kept in it's labeled class in the artificial data set. If a data point crosses the boundary of its respective class, then a K-NN based verification is applied. If 3 out of the nearest 5 neighbors belongs to the class to which the data point is supposed to belong, then the data point is kept in the artificially augmented data set. If both the conditions above fail, then the respective data point's class label is changed so that it belongs to the class whose properties it corresponds to better. Steps are repeated for all the artificial data points generated, in sequence.

2. **By estimating an empirical distribution from the data:** Proposed statistical foundations shall be laid to analyze and handle the challenges in the data set. This has not been attempted before in the available literature. Simulations of the augmentation of the samples in the data set and the results of the ML methods shall be tried on the artificially augmented data samples. The purpose of this exercise is to emulate the natural process of discovering new exoplanets and trying to classify them.
   Generating Data by Analyzing the Distribution of Existing Data Empirically: Window Estimation Approach: In this method of synthesizing data samples, the density of the data distribution is approximated by a numeric mathematical model, instead of relying on an established analytical model (such as Poisson, or Gaussian distributions). As the sample distribution here is sporadic, the density function itself should be approximated. The process is termed Kernel Density Estimation (KDE). KDE, as a non-parametric technique, requires no assumptions on the structure of the data and further, with slight alterations to the kernel function, may also be extended to multivariate random variables.

   Estimating Density:: Let $X = x_1, x_2, \ldots, x_n$ be a sequence of independent and identically distributed multivariate random variables having d dimensions. The window function used is a variation of the uniform kernel defined on the set R^d as follows:

   $$\phi(u) = \begin{cases} 1 & u_j \leq \frac{1}{2} \quad \forall j \in \{1, 2, \ldots, d\} \\ 0 & otherwise \end{cases} \tag{1}$$

   Additionally, another parameter, the edge length vector $h = \{h_1, h_2, \ldots h_d\}$, is defined, where each component of h is set on a heuristic that considers the values of the corresponding feature in the original data. If f_j is

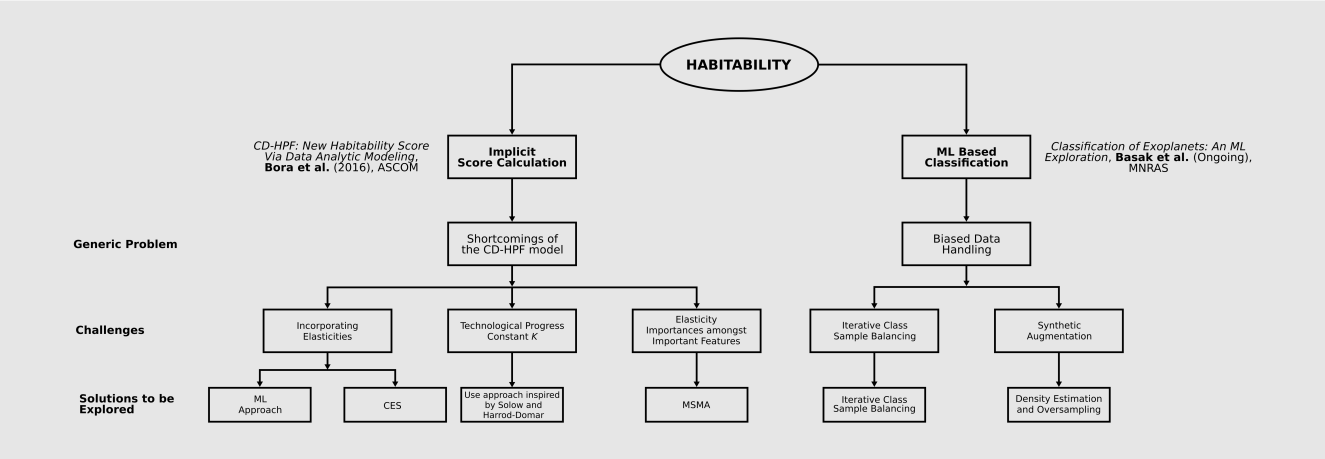
HABITABILITY
CD-HPF: New Habitability Score Via Data Analytic Modeling, Bora et al. (2016), ASCOM
Implicit Score Calculation
ML Based Classification
Classification of Exoplanets: An ML Exploration, Basak et al. (Ongoing), MNRAS
Generic Problem
Shortcomings of the CD-HPF model
Biased Data Handling
Challenges
Incorporating Elasticities
Technological Progress Constant K
Elasticity Importances amongst Important Features
Iterative Class Sample Balancing
Synthetic Augmentation
Solutions to be Explored
ML Approach
CES
Use approach inspired by Solow and Harrod-Domar
MSMA
Iterative Class Sample Balancing
Density Estimation and Oversampling

Figure 1: METHODOLOGY FLOW: A magnified image is available here: https://drive.google.com/file/d/0B1yuFIvTaFbhUWY4enNQUlU2eW8/view?usp=sharing

the column vector representing some feature $j \in X$ and

$$\begin{aligned} l_j &= min\{(a-b)^2 \quad \forall\ a,b \in f_j\} \\ u_j &= max\{(a-b)^2 \quad \forall\ a,b \in f_j\}, \end{aligned} \tag{2}$$

the edge length h_j is given by,

$$h_j = c\left(\frac{u_j + 2l_j}{3}\right) \tag{3}$$

where c is a scale factor. Let $x' \in R^d$ be a random variable at which the density needs to be estimated. For the estimate, another vector u is generated whose elements are given by: $u_j = \frac{{x_j}' - x_{ij}}{h_j} \qquad \forall j \in \{1, 2, \ldots, d\}$. The density estimate is then given by the equation: $p(x') = \frac{1}{n\prod_{i=1}^{d} h_i} \sum_{i=1}^{n} \phi(u)$

Binomial distribution based confidence splitting criteria in the Random Forest classifier approach: The binomially distributed probability of correct classification of an entity may be used as a node-splitting criteria in the constituent DT of an RF. From the cumulative binomial distribution function, the probability of k or more entities of class i occurring in a partition A of n observations with probability greater than or equal to p is given by the binomial random variable as in Equation: $X(n,p) = P[X(n,p_i) \geq k] = 1 - B(k; n, p_i)$. As the value of $X(n,p)$ tends to zero, the probability of the partition A being a pure node, with entities belonging to only class i, increases. However, an extremely low value of $X(n,p)$ may lead to an over-fitting of of data, in turn reducing classification accuracy. A way to prevent this is to use a *confidence threshold*: the corresponding partition or node is considered to be a *pure* node if the value of $X(n,p)$ exceeds a certain threshold. Let c be the number of classes in the data and N be the number of *outputs*, or *branches* from a particular node j. If n_j is the number of entities in the respective node, k_i the number of entities of class i, and p_i the minimum probability of occurrence of k_i entities in a child node, then the model for the confidence based node splitting criteria as used by the authors may be formulated as equation: $var = \prod_{j=1}^{N} min\{1 - B(k_{ij}; n_j, p_j)\}$, where $\mathcal{I} = 0$ if $var <$ confidence threshold, else is var; subject to the conditions, $c \geq 1$, $p = [0,1]$, confidence threshold $= [0,1)$, $k_{ij} \leq n_j$, $i = \{1, 2, ..., c\}$, $j = \{1, 2, ..., N\}$. Here, the i subscript represents the class of data, and the j subscript represents the output branch. So, k_{ij} represents the number of expected entities of class i in the child node j.

1.1 Perturbed Econometric modeling and impact on Habitability: Handling the eccentricity feature

The major limitation of using the Cobb-Douglas model to compute the habitability of an exoplanet by the means of the CD-HPF function is due to the multiplicative nature of the model. In case the value of any feature is zero, the entire score results in zero, which is not a valid score. In order to account for zero-valued features (and by incorporating some prior knowledge), we propose a new model called the Perturbed Augmented CD-HPF (PACD-HPF). We need to prove the following theorem:

Theorem: If global maxima for CDHS, i.e.

$$\log(Y) = \frac{1}{1 - \sum\limits_{i=1}^{n} \alpha_i} \log \left\{ k \prod_{i=1}^{n} \left(\frac{x_i p}{w_i} \right)^{\alpha_i} \right\} \tag{4}$$

holds [1], then the same condition for the global maxima will continue to hold if an additional input parameter is inserted in the habitability function CD-HPF, i.e., if

$$\log(Y_{\text{new}}) = \frac{1}{1 - \sum\limits_{i=1}^{n+1} \alpha_i} \log \left\{ k \prod_{i=1}^{n+1} \left(\frac{x_i p}{w_i} \right)^{\alpha_i} \right\}$$

holds as well. Further, it follows that the elasticity condition for DRS for $n+1$ parameters is true, i.e. $1-\sum\limits_{i=1}^{m+1} \alpha_i > 0$, if the elasticity condition for DRS for n parameters, i.e. $1-\sum\limits_{i=1}^{m} \alpha_i > 0$, holds. Through this model, we are interested in accounting for two new features: the *orbital velocity* of an exoplanet, and the *eccentricity* of an exoplanet. In a more specific sense, the problem here is that the feature *eccentricity* has a lot of zero-values in the catalog that we have used. However, it is common knowledge that the orbits of most planets are not perfectly circular, for the value of eccentricity to be zero. Abel Méndez and his team, who have put together the PHL-EC catalog, have used the value of zero for eccentricity in cases where it was not possible to be estimated or observed. However, eccentricity is an important feature and must be taken into account. We do this by first perturbing the value of eccentricity. Let E denote the value of eccentricity after perturbation, and E_0 be the original value, which could be equal to zero. Then, the perturbation is formulated as: $E = aE_0 + b$, where a is a scaling coefficient, and b is the perturbation factor. We incorporate E and V_o (orbital velocity) in the Cobb Douglas model as: $\mathbb{Y} = f(R, D, T_s, V_e, V_o, E) = (R)^{\alpha_1} \cdot (D)^{\alpha_2} \cdot (T_s)^{\alpha_3} \cdot (V_e)^{\alpha_4} \cdot (V_o)^{\alpha_5} \cdot (aE_0 + b)^{\alpha_6}$

In addition to maximization, we will find the combination of all α_i, such that the separation between the classes of potentially habitable and non-habitable planets is maximum. This can be done by incorporating the ideas of support vector machines for margin maximization. We formulate the complete problem as maximizing the margin between the two classes as: $\frac{1}{2} w \cdot w + C \sum_i^n \epsilon_i$, which is the typical formulation of soft-margin SVM's, subject to the constraints: $y_i(\phi(x_i) \cdot w + b) \geq 1 - \epsilon_i$, where ϕ is a multi-dimensional kernel defined by the Cobb-Douglas function.

2 Brain-Computer Interfacing

I got interested in Brain-Computer interfacing around December 2014. I tried to develop a BCI using the NeuroSky MindWave Mobile, which is a small headset and can be connected via bluetooth. Till January 2016, I worked on different methods to process the data and perform a binary classification. The idea was to collect two types of signals from test subjects (corresponding to a YES or NO) and classify them. The final classification was used with a speller-machine. The rationale for this project was to develop a system which makes the communication of disabled individuals, who are incapable of muscular coordination (conditions such as ALS).

My work on BCI has been published as a book chapter in Handbook of Research on Applied Cybernetics and Systems Science (Advances in Computational Intelligence and Robotics) (IGI Global)

3 Indic Tools

The Indic Tools project aimed at annotating images of text documents that were originally written in Indian languages, such as Hindi, Tamil, Sanskrit, etc. by employing pattern matching of words over the same image file as well as other image files. Upon annotating a word, the program looked for similar patterns as the selected word in the image and tells the user that the matched patterns could possibly have the same meaning. This would make annotation in documents easier. The objective of the project was to provide an interface for users to share

[1] Remark:The habitability score CDHS is computed using four parameters: R, D,T_s and V_e. If a new parameter from the PHL-EC needs to be added to the CD-HPF, it is important to know if the conditions of global maxima for habitability still holds. The above theorem validates our superposition conclusively.

information that isn't directly available in English in images of text.

My contributions to this chiefly were:

1. Binarization of relatively noisy images in a sequence of steps, including histogram equalization, edge detection and adaptive thresholding.
2. Words are segmentation based on contours, and elimination of noise and/or outliers statistically
3. Skew angle corrected by taking a random sample from the set of detected words, and computing the individual angles of each word. The angles are approximated to the nearest whole number, and the median of these angles is taken as the skew angle of the image file.

I pursued this project under Dr. Kanchi Gopinath, Department of Computer Science and Automation, IISc. My involvement in this project lasted between April and September 2015, after which, it was brought to a conclusion.

4 Alternate Interface for Electronic Drum Kits Based On Computer Vision

This was the first research project that I pursued. We began it in September 2014, and brought it to a conclusion around March 2015.

In this project, we explored computer vision for developing an alternate interface for electronic drumkits. The idea was to connect a camera to a computer and draw circles on a flat surface, each circle representing a drum. Then the user would have to select a sound corresponding to each circle. The drum sticks were also connected to a computer, with an intermediate force sensor to estimate how hard a drum was struck. The combination of which circle's the tip of the drumstick came inside with the force of the drumstick led to a specific sound being played. The rationale behind this was to improve portability for musicians. Although we never completed the GUI of the application, a proof-of-concept model was ready, which we presented at the RISE conference in PESIT Bangalore South Campus.

The technologies we explored overall for this included multi threaded programming, computer vision, audio playback, and programmable hardware. The paper is available at: http://pesitsouth.pes.edu/rise/papers/csis/CSIS004.pdf