## Nova data imputation

**Data set statistics:**

* Totally there are 103 nova files, out of which currently I have processed 88 of them which are small in size.
* There are 27 features in each file namely 'JD', 'Star.Name', 'Band', 'Observer.Code', 'Comment.Code.s.', 'Comments', 'Charts', 'Observer.Affiliation', 'Validation.Flag', 'Measurement. Method', 'Grouping. Method', 'ADS. Reference', 'Digitizer', 'Credit', 'HJD', 'HQuncertainty', 'Transfomed', 'Star Name'
* I have selected following features for the work as they have more amount complete rows across the dataset. Also rest of the features makes little sense with respect to our task in hand.
  + Cmag ,Kmag, Airmass, Uncertainty, Comp.Star.1, Comp.Star.2, Magnitude
* The overall data characteristics for each of the feature is as shown in attached data\_stats.xsl
* As we can see the data is very sparse in fact in multiple cases that has only nulls in many columns.
* However Magnitude, Comp star 1 & 2 features are completely filled and usable.

**Imputation method**

* To begin with we shortlisted following methods for imputation purposes namely.
  + Chained equations[1]
  + Multivariate normal distributions approach[2]
  + Random forest with out of bag error metric[3]
  + Fischer scoring[4]
  + Nearest neighbor with a low rank approximation using SVD[5]
* Considering dataset statistics since in most datasets some of the rows of Magnitude, Comp star 1 & 2 are available we decided to deduce the missing values using chained equations.

**Method**

The method used currently is known as multiple imputations using chained equations where we impute same data multiple times, as opposed to single imputation [2-5], there by the variance of the imputation is minimized. More specifically we used chained equations approach for multiple imputation, where we have X1, X2….Xk features. If X1 has missing values, then it will be regressed on other variables X2 to Xk. The missing values in X1 will be then replaced by predictive values obtained. The process of mice and selection criteria on imputation cycles are detailed in [1]. For our work we used cart [6] based imputation procedure with multiple imputation, where for each imputation cycle and each missing value ymis we execute following steps

1. Fit a classification or regression tree by recursive partitioning;
2. For each ymis, find the terminal node they end up according to the fitted tree;
3. Make a random draw among the member in the node, and take the observed value from that draw as the imputation.

Currently for this imputation process we selected following values for various hyper parameters.

|  |  |
| --- | --- |
| **Hyper parameters** | **Values** |
| Imputation cycles | 50 |
| Method | CART |
| Number of members for random draw | 5 |
| Split complexity (min) | 0.0004 |

The source code, data can be found in <https://github.com/manikandan-ravikiran/Nova>

Citation:

1. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011;20(1):40–49.
2. Schafer JL. , Analysis of Incomplete Multivariate Data , 1997London, United Kingdom Chapman & Hall Ltd
3. Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani (2013). An Introduction to Statistical Learning. Springer. pp. 316–321.
4. Briers, Mark and Winston Churchill. “Improved Monte Carlo Methods for State-Space Models.” (2007).
5. Batista, Gustavo E. A. P. A. and Maria Carolina Monard. “A Study of K-Nearest Neighbour as an Imputation Method.” HIS (2002).
6. Doove, L.L., van Buuren, S., Dusseldorp, E. (2014), Recursive partitioning for missing data imputation in the presence of interaction Effects. Computational Statistics \& Data Analysis, 72, 92-104.