Feature Selection Strategies for Classifying High Dimensional Astronomical Data Sets

Ciro Donalek.et al.

Yule Wang

May 29,2024

Content

- 1. Introduction
- 2. Datasets
- 3. Feature Selection Strategies
- 4. Classifier Design
- 5. Results
- 6. Conclusions

1. Introduction

- Due to the recent surge in astronomical data, in many cases, this plethora of data need to be processed in quasi real-time fashion in order to the potential usable value from the data as quickly and efficiently as possible.

- An automated reliable and robust classification of transient and variable sources is of a critical importance for the effective follow-up of the present and future synoptic sky surveys, This is because when they are discovered, if not classified, all transients appear to be the same.

- But most systems today rely on a delayed human judgment in decision making and follow up of events and this “manual” approach will simply not scale to the next generation of surveys.

- A major problem associated with pattern recognition in high-dimensional data sets is **the limited of dimensionality**, but this can be resolved by **selecting only a subset of features that are rich in discriminatory power.**

- So, the research objective of this paper is to **design a high-dimensional feature dataset with simple descriptors and a high-performance classifier**.

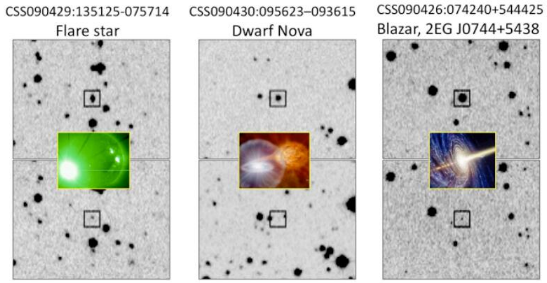


Fig. 1. Top row shows objects which appear much brighter that night, relative to the baseline images obtained earlier (bottom row). On this basis alone, the three transients are physically indistinguishable. Subsequent follow-ups show them to be three vastly different types of phenomena.

2.Datasets

- CRTS(Catalina Real-Time Transient Survey)

- Kepler Data

2.Datasets

- Each astronomical object in represented through its light curve that can be sparse and uneven sampled, with tremendous variations also in errors, number of points, missing values, etc., making comparison between them as well as training the classifiers difficult.

- So the authors using the Caltech Time Series Characterization Service (CTSCS) vectors of such features derived from the light curve of known classes of objects were used as training and test sets in our classification system.

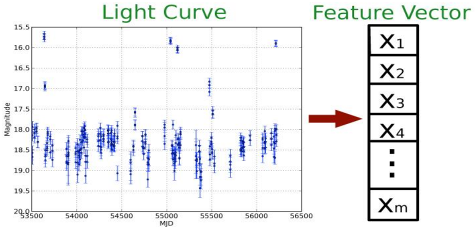


Fig. 2. Through the CTSCS a set of statistical descriptors are extracted from each light curve forming a feature vector. The plot shows a light curve of a Cataclysmic Variable from CRTS.

- Catalina Real-Time Transient Survey (CRTS) covers the total area of ~ 33,000 deg^2 , down to ~ 19 - 21 mag per exposure, with time baselines from 10 min to 8 years, and growing, To date, the survey has discovered many unique transient objects (including cataclysmic variables (CVs), supernova, quasars, etc.). It has also released over 500 million light curve data points. Notably, the observational data obtained from the survey is open source.

- The speciality of the Kepler observations are the dense lightcurves required for various planet detection methods and because of the relatively small area it covers (115 square degrees).

3.Feature Selection Strategies

- Feature selection algorithms can be roughly grouped in to two categories: filter methods and wrapper methods.

- The filter methods does not involving a specific learning algorithm, where as the wrapper method does involve a specific learning algorithm because it uses a chosen algorithm to evaluate candidate feature subsets.

- The author introduces five different algorithms and assesses each of these algorithms in subsequent.

3.Feature Selection Strategies

- FRA: It usually applied in data pre-processing as a feature subset selection method. The key idea of the FRA Algorithm is to estimate the quality of attributes according to how well their values distinguish between instances.
- FDR: Can be used to rank a number of features with respect to their class discriminatory power and can be independent of the type of the underlying class distribution.
- CFS: It is a wrapper method which selects features that have low redundancy and is strongly predictive of a class.
- FCBF: It is a supervised filter based feature selection algorithm. This method is similar to the CFS algorithm in the sense that it also uses feature-class correlations and symmetrical uncertainty as measures of goodness.
- MCFS: It is basically an unsupervised feature selection method based on the spectral analysis of the data. The specialty of this algorithm is that, even though inherently unsupervised, it can be used in supervised as well as semi-supervised modes.

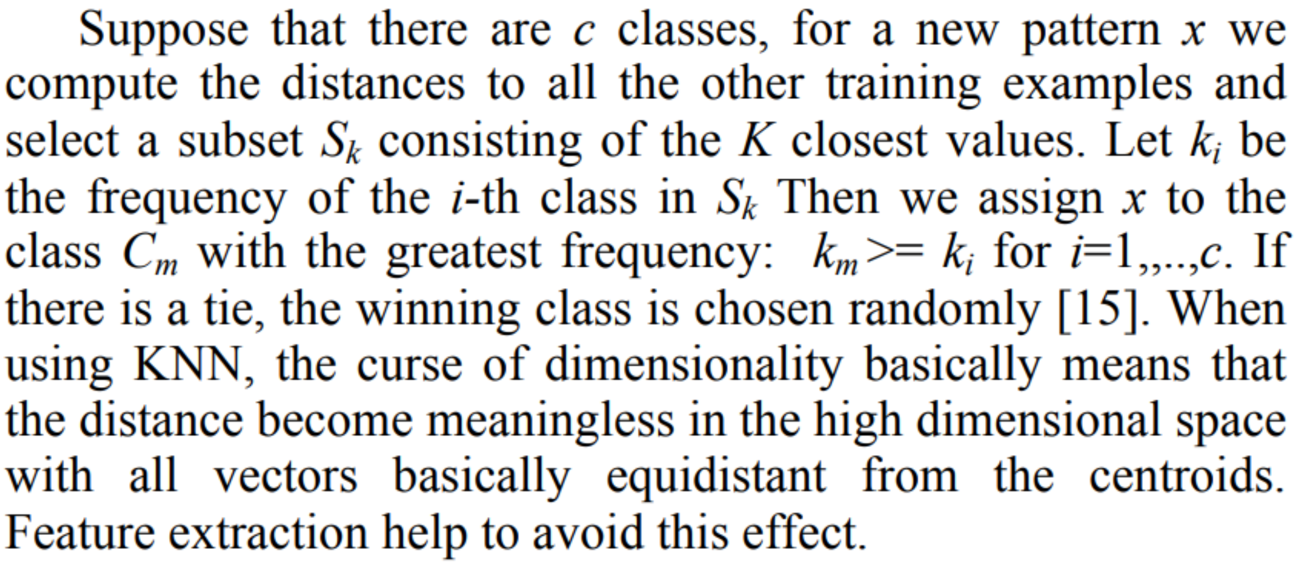
4.Classifier Design

- Ensembles of K-nearest-neighbor (KNN)

- Ensembles of Decision Trees

4.Classifier Design

- The K-nearest-neighbor (KNN) is a very simple method for classification， the method is often used as a benchmark method.



Suppose that there are c classes, for a new pattern x we compute the distances to all the other training examples and select a subset S_k consisting of the K closest values. Let k_i be the frequency of the i-th class in S_k Then we assign x to the class C_m with the greatest frequency: $k_m >= k_i$ for i=1,,..,c. If there is a tie, the winning class is chosen randomly [15]. When using KNN, the curse of dimensionality basically means that the distance become meaningless in the high dimensional space with all vectors basically equidistant from the centroids. Feature extraction help to avoid this effect.

4.Classifier Design

- Decision Trees (DT, aka Classification Trees) are one of the most used data mining, non-linear classifier. In a decision tree the feature space is split into unique regions, corresponding to the classes. Each internal node denotes a test on an attribute, each branch represents the outcome of the test and each leaf holds a class label.

- The method using all the features available, and then on the substes generated by the feature selectio algorithm described above. Each tree was built using the Gini Diversity Index (gdi) as criterion for choosing the split; the splitting stops when there is no further gain that can be made.

5. Experiments

- Supernova Classification

- WuMa vs RR Lyrae in CRTS

- Classifying objects in Kepler Data

5. Experiments

- The authors found that an hierarchical approach led to better results. They used some astrophysically motivated major features to separate different groups of classes (see Fig. 3), and then proceeding down the classification hierarchy each node uses those classifiers that are demonstrated to work best for that particular task.

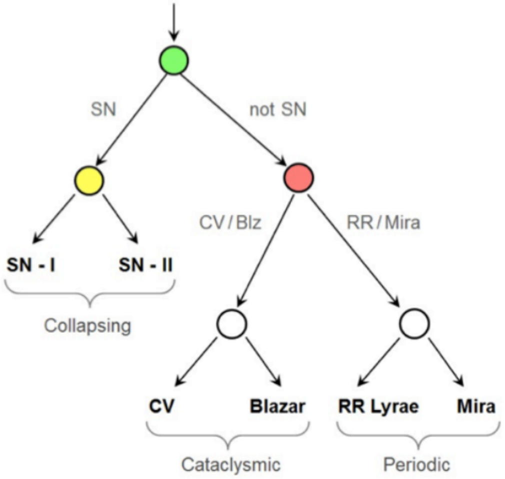


Fig. 3. Hierarchical approach: some astrophysically motivated major features are used to separate different groups of classes. Proceeding down the classification hierarchy each node uses those classifiers that are demonstrated to work best for that particular task.

5. Experiments

- For the binary classification problem of supernovae, the goal is to reliably assign each object to one of two mutually exclusive categories: supernova or non-supernova. To this end, the authors extracted objects from six different categories from the light curve data of CRTS, and then used CTSCS to extract twenty features from each light curve data as the training set. The results obtained are displayed in Table I.
- The Figure 4 shows the change in misclassification errors when training the decision tree with different numbers of features; by using the top six parameters ranked by the FRA algorithm to train the decision tree, the best results were obtained, which significantly improved the accuracy for both KNN and decision tree classifiers.

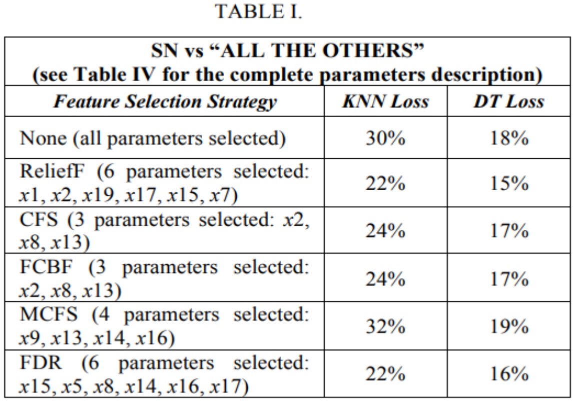


TABLE I.

SN vs “ALL THE OTHERS” (see Table IV for the complete parameters description)		
Feature Selection Strategy	***KNN Loss***	***DT Loss***
None (all parameters selected)	30%	18%
ReliefF (6 parameters selected: $x1, x2, x19, x17, x15, x7$)	22%	15%
CFS (3 parameters selected: $x2, x8, x13$)	24%	17%
FCBF (3 parameters selected: $x2, x8, x13$)	24%	17%
MCFS (4 parameters selected: $x9, x13, x14, x16$)	32%	19%
FDR (6 parameters selected: $x15, x5, x8, x14, x16, x17$)	22%	16%

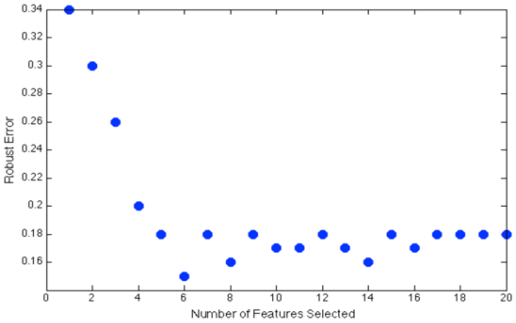


Fig. 4. The plot shows how the misclassification error change based on the number of features used to train the DT. The features were ranked using the ReliefF algorithm. Best result was achieved using the six most important features found by the algorithm.

5. Experiments

- For the WuMa vs RR Lyrae in CRTS problem, the author extracted light curves for 482 RR Lyrae and 463 W UMa from CRTS, and used the CTSCS to extract 60 periodic and non-periodic features for each object,and using the five higher ranked parameters according the FRA algorithm.
- Table II show completeness and contamination for both classes, achieved using all the parameters and the selected subset. Not only using the best five parameters according to the ReliefF algorithm decreased dramatically the computational time but also led to a higher completeness and lower contamination.
- The parameters automatically selected by this procedure, essentially represent the period-amplitude relationship illustrated in Fig. 5 which is used to differentiate between subclasses of RR Lyrae.

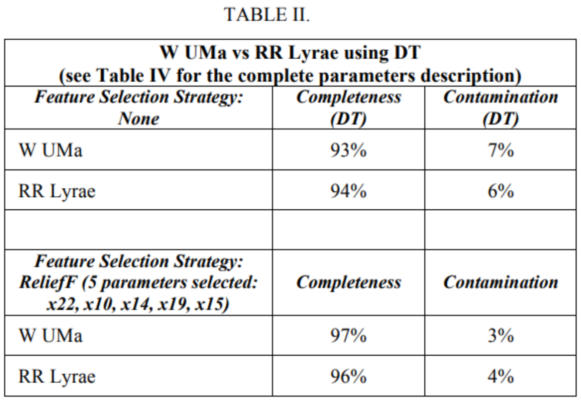


TABLE II.

W UMa vs RR Lyrae using DT **(see Table IV for the complete parameters description)**		
Feature Selection Strategy: None	***Completeness (DT)***	***Contamination (DT)***
W UMa	93%	7%
RR Lyrae	94%	6%
Feature Selection Strategy: ReliefF (5 parameters selected: x22, x10, x14, x19, x15)	***Completeness***	***Contamination***
W UMa	97%	3%
RR Lyrae	96%	4%

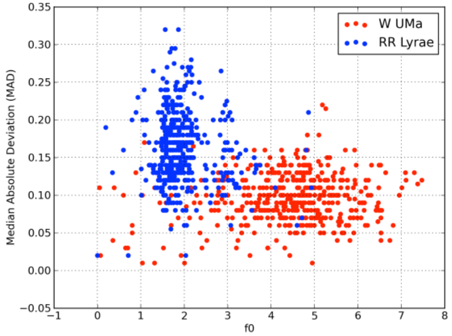
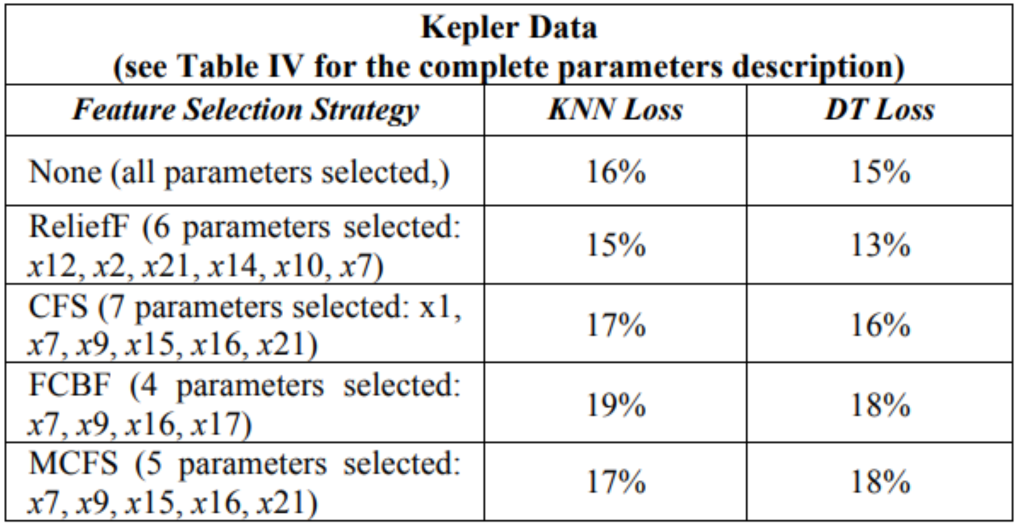


Fig. 5. It is interesting to note that the best two parameters automatically selected by the ReliefF algorithm (f0, mad), essentially represent the period-amplitude relationship [7] and give a good separation between the two classes.

5. Experiments

- In this problem,the subsets authors have considered here include ~20 thousands objects that Kepler has observed, also contains many false positives which are also of interest to authors. That's why a reliable classification of these objects is needed.
- The authors ran four feature selection strategies to a 3-class problem, and then test the subsets generated by them using both KNN and DT. Table III show the results achieved; the best overall classification was reached using a DT and the six best parameters found by the FRA algorithm.

TABLE III.



Kepler Data
(see Table IV for the complete parameters description)

Feature Selection Strategy	***KNN Loss***	***DT Loss***
None (all parameters selected,)	16%	15%
ReliefF (6 parameters selected: *x*12, *x*2, *x*21, *x*14, *x*10, *x*7)	15%	13%
CFS (7 parameters selected: x1, *x*7, *x*9, *x*15, *x*16, *x*21)	17%	16%
FCBF (4 parameters selected: *x*7, *x*9, *x*16, *x*17)	19%	18%
MCFS (5 parameters selected: *x*7, *x*9, *x*15, *x*16, *x*21)	17%	18%

6. Conclusions

- In this section,the authors summarizes and emphasizes the importance of feature selection in analyzing multiparametric astronomical datasets to reduce the dimensionality of the input space, Through three specific experimental cases show the advantages of employing feature extraction routines instead of utilizing all available features are showcased.