Flare Forecasting

Goal: To predict whether an active region will flare, or not, within a given time interval

Machine Learning:
Field of computer science that develops algorithms to learn a specific task without being explicitly programmed for it

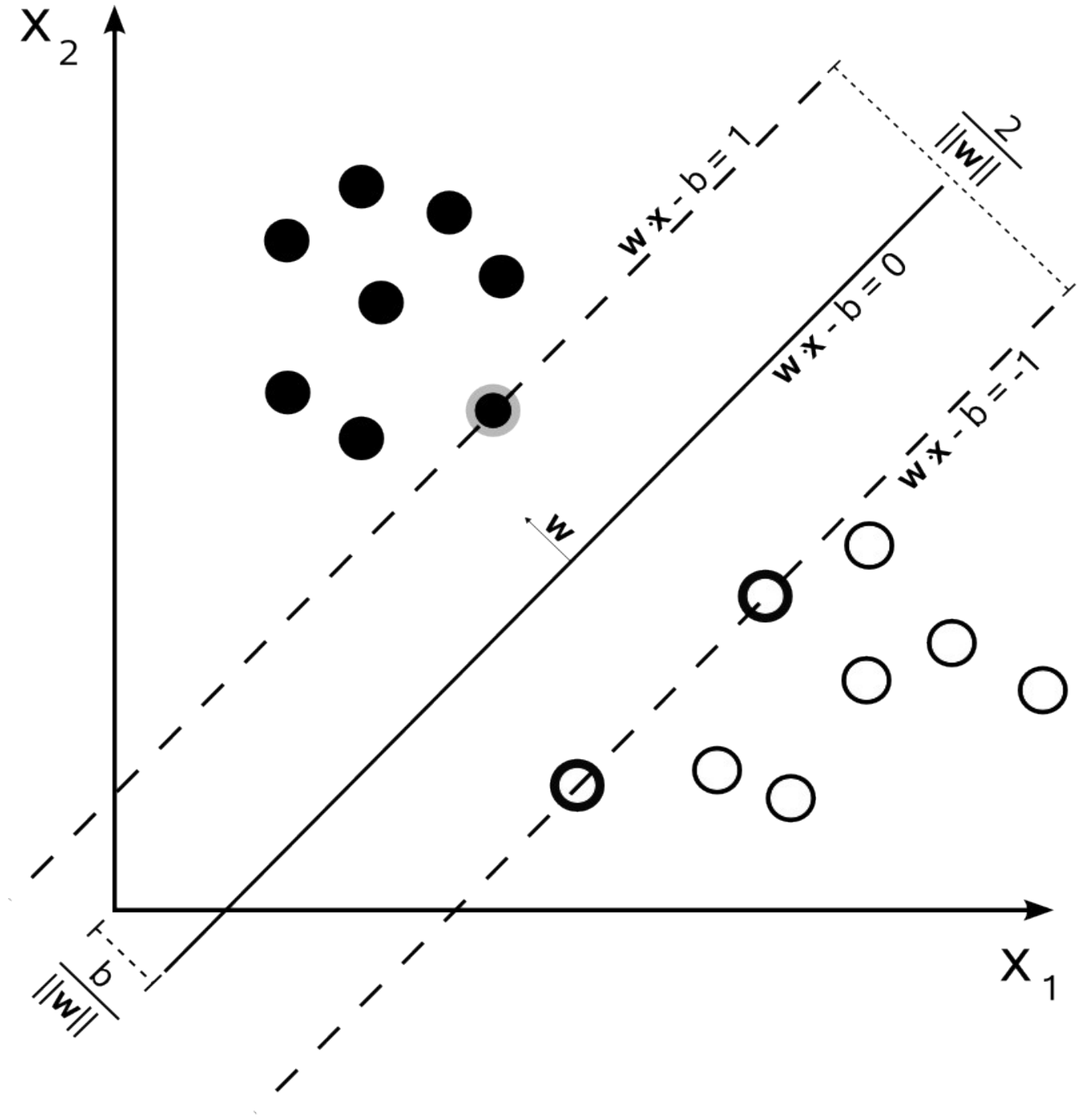
Li, Wang, Cui, & Du (2007): support vector machine + k-nearest neighbour
Colak & Qahwaji (2007): neural network
Yu, Huang, Wang, & Cui (2009): decision tree
Song, Tan, Jing, et al. (2009): logistic regression
Yuan, Shih, Jing, & Wang (2010): support vector machine
Ahmed, Qahwaji, Colak, et al. (2013): cascade-correlation neural network

Non Machine Learning:

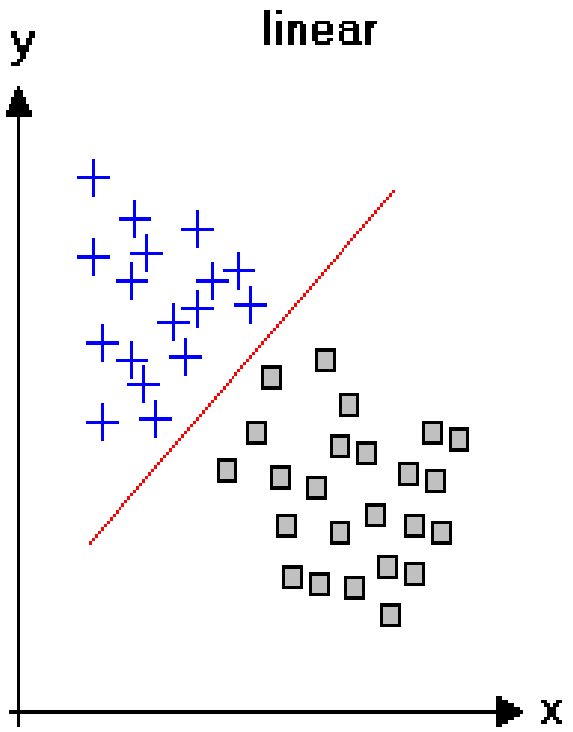
Wheatland (2005): statistical approach based on active region history
Barnes & Leka (2003,2008): discriminant analysis
Mason & Hoeksema (2010): epoch analysis
Falconer, Moore, Barghouty, et al. (2012): prior flaring + free magnetic energy

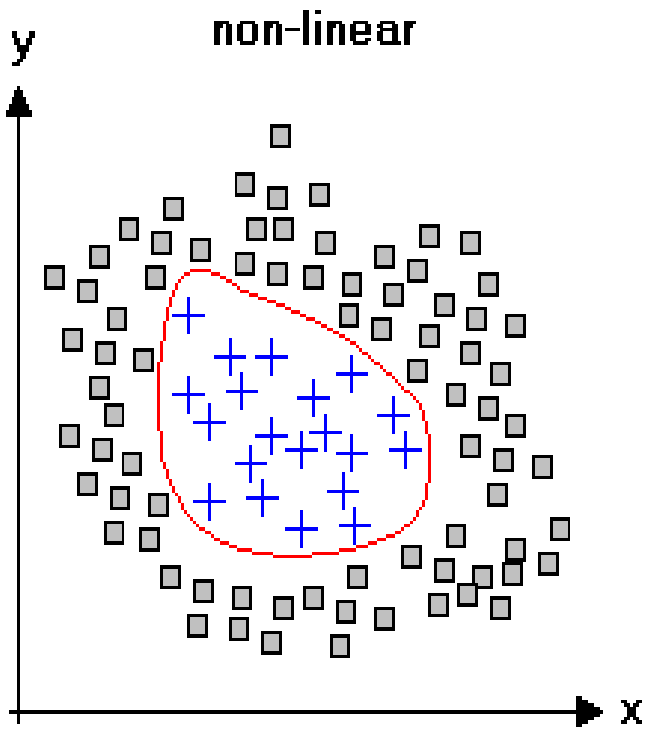
Here, we use support vector machine with HMI vector magnetograms (also tried: neural network, decision trees, logistic regression, and Adaboost). We used 303 flaring ARs and 5000 non-flaring ones

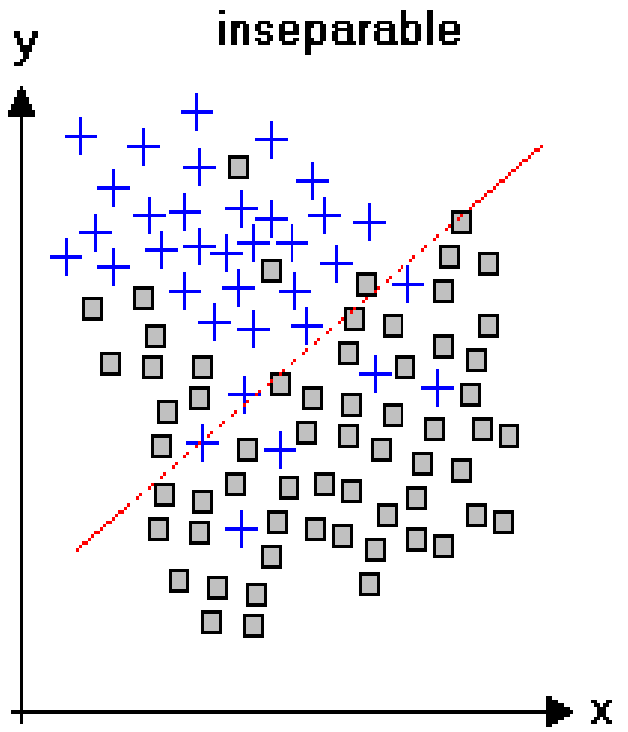
Support Vector Machine



- a binary classifier known to be superior to neural networks
- soft margin version proposed by Cortes & Vapnik (1995) works on non-linearly separable features
- tries to separate the features by an hyperplane with the largest margin while minimizing mis-classifications
- cost function=1/2 ||w||^2 + C*Σε
- decision function can be non-linear through the use of kernels
- we use the Scikit-Learn module in Python (based on libsvm) commonly used in the literature
- we use a Gaussian kernel
=> 2 main parameters to train the SVM
- example dataset randomly separated in training and testing sets (70-30%)


linear
y
x


non-linear
y
x


inseparable
y
x

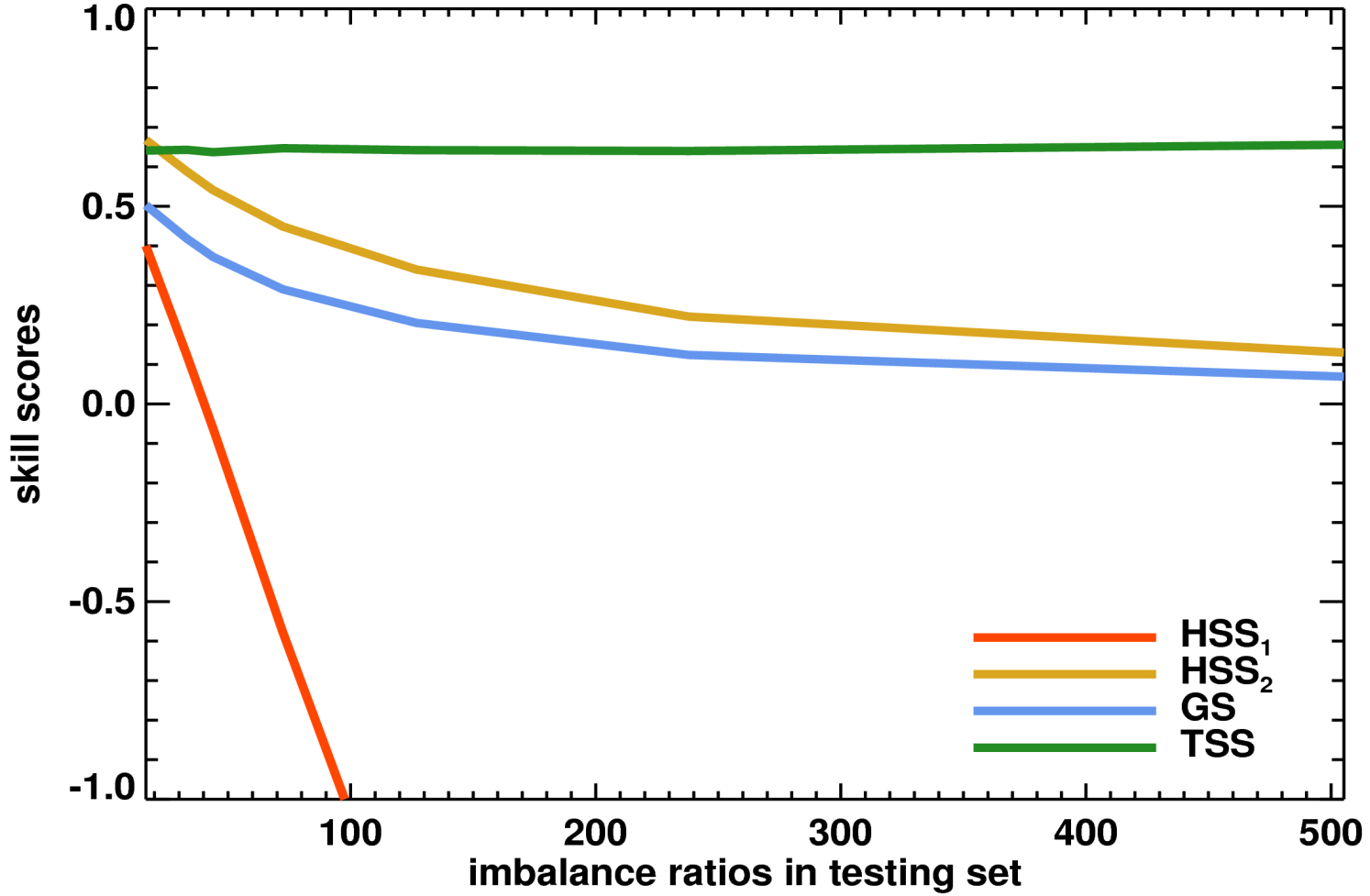
Major Issue For Flare Prediction: Class Imbalance

- Solar flare forecasting is affected by strong class imbalance: many more negative examples N (non-flaring ARs) than positive ones P (e.g., Mason & Hoeksema had N/P=260 in their Table 2; Ahmed et al. have ~17; Barnes & Leka have ~10)
- different ways to deal with it: here, we assign different cost functions to the two classes

- imbalance impacts performance metrics used by various groups:
To measure performance, we use a contingency table: TP, TN, FP, FN
Accuracy=(TP+TN)/(P+N)
Precision=ability of the classifier not to label as positive a negative example=TP/(TP+FP)
Recall (aka sensitivity, POD...)=ability of the classifier to find all of the positive examples=TP/P
Skill scores measure the performance compared to a benchmark:
Heidke Skill Score (Barnes & Leka, 2008) HSS_1: correct predictions compared to always predicting negative
Heidke Skill Score (Balch 2008) HSS_2: correct predictions compared to random predictions
Gilbert Skill Score (Mason & Hoeksema, 2010) GS: correct positive predictions (TP) compared to correct positives from random predictions

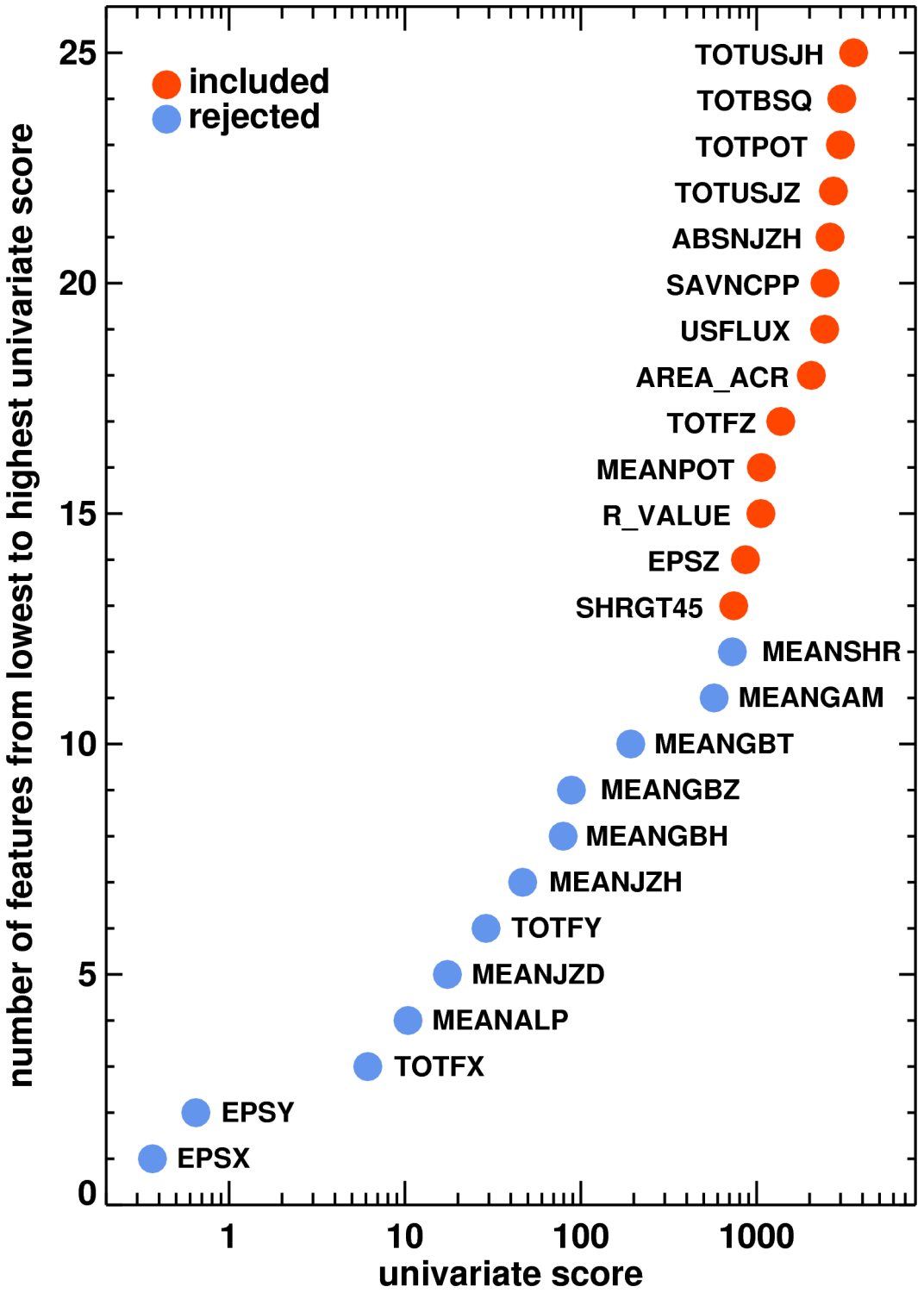
- best to use **True Skill Statistic** (Bloomfield et al., 2012) TSS (aka Hanssen-Kuiper skill score), is: recall minus false alarm rate =TP/P-FP/N (or recall+specificity-1). Widely used to test the performance of weather forecasts (McBride & Ebert, 2000)

Sensitivity of Skill Scores to N/P Ratio



Only TSS independent of imbalance ratio (Woodcock, 1976; Bloomfield et al. 2012) => TSS should be the preferred performance metric when comparing results of groups with different N/P ratios

NB: HSS_1 is the least useful skill score (outside of accuracy) in case of strong imbalance



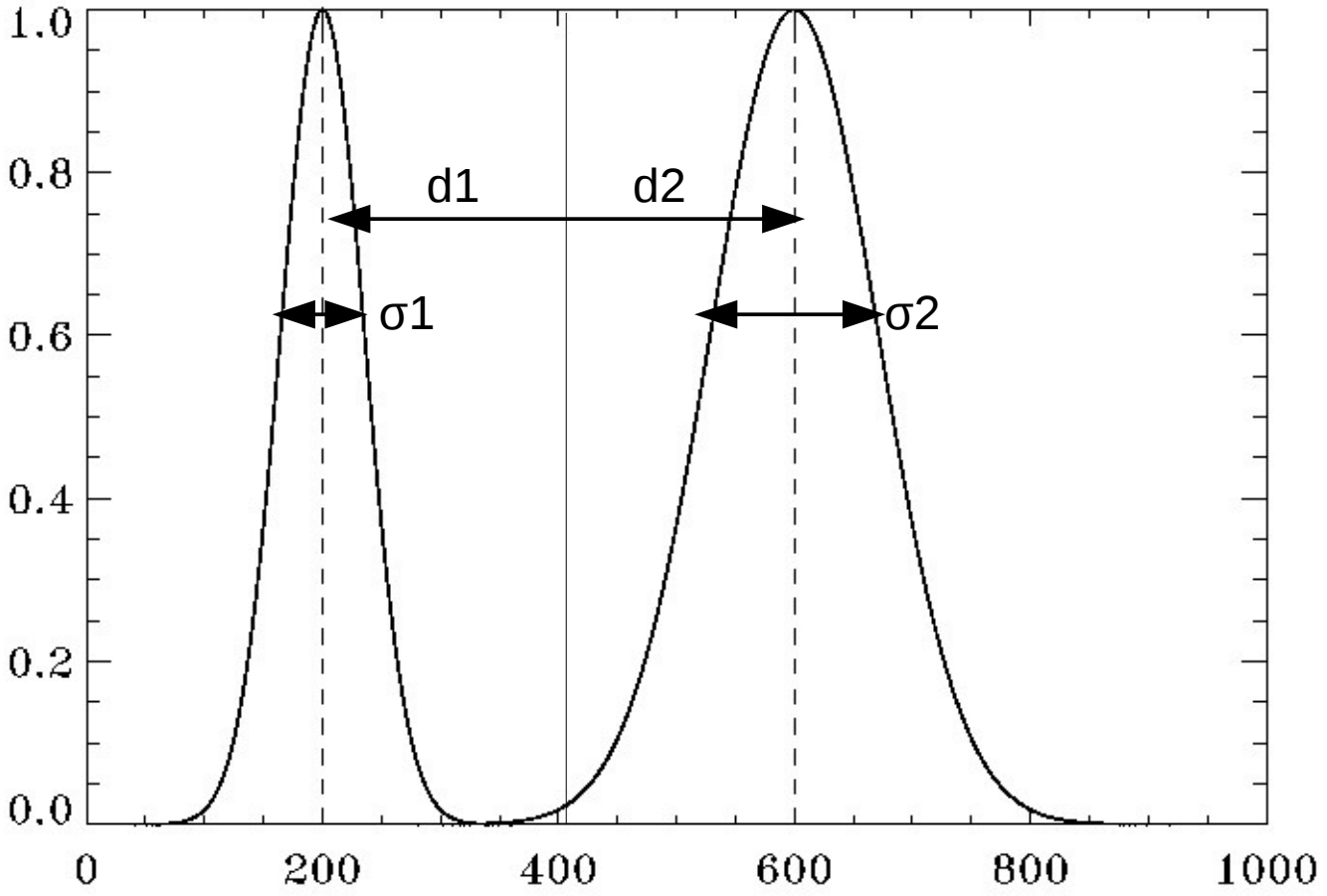
Feature Selection

- including too many features does not improve the performance of SVM (and may lower the performance of other ML algorithms) => we try to optimize the number of features
- we use univariate feature selection based on the Fisher ranking score
- easy to implement and useful, but ignores any dependence between features (compute correlation coefficients)
- score has significant error bars (depends on the number of examples and the ones selected)
- we retain only 13 features

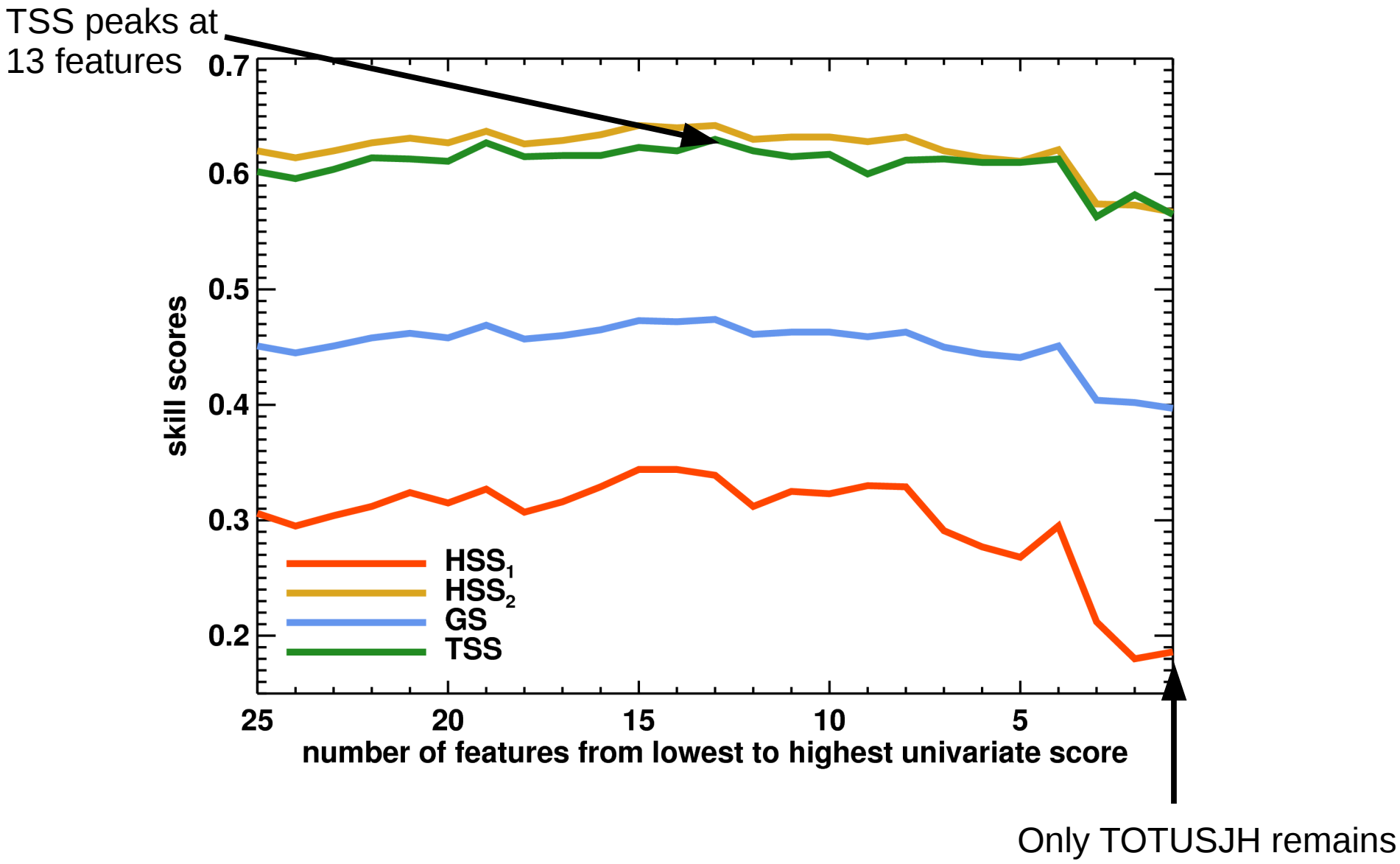
- New feature-selection algorithms tested based on Shannon's entropy and maximum-relevance-minimum-redundancy criteria (multivariate feature selection)

Fisher Ranking Score

Fisher score = $(d1^2+d2^2)/(\sigma1^2+\sigma2^2)$ (based on F-statistic)



Skill score as a function of number of features



Results

Metric			Mason	Ahmed	Ahmed	Barnes	Bloomfield	Yu	Song
Time interval (no flare)	48h	24h	6h	48h	24h	24h	24h	48h	24h
class-imbalance ratio	16.5	16.5	260	15.85	16.58	9.92	26.5	NA	2.23
Accuracy	0.973±0.004	0.963±0.004	0.694	0.975	0.963	0.922	0.830	0.825	0.873
Precision (positive)	0.779±0.047	0.661±0.034	0.008	0.877	0.740	NA	0.146	0.831	0.917
Precision (negative)	0.984±0.003	0.980±0.002	0.998	0.980	0.972	NA	NA	NA	0.860
Recall (positive)	0.732±0.047	0.641±0.045	0.617	0.677	0.523	NA	0.704	0.817	0.647
Recall (negative)	0.987±0.003	0.982±0.004	0.695	0.994	0.989	NA	NA	NA	0.974
f1 (positive)	0.753±0.032	0.661±0.034	0.015	0.764	0.613	NA	0.242	NA	0.758
f1 (negative)	0.986±0.002	0.980±0.002	0.819	0.987	0.989	NA	NA	NA	0.913
$IISS_1$	0.520⊥0.064	0.342⊥0.076	-78.9	0.581	0.339	0.153	NA	NA	0.588
HSS_2	0.739±0.033	0.641±0.035	0.008	0.751	0.594	NA	0.190	0.650	0.676
Gilbert skill score	0.587±0.042	0.473±0.039	0.004	0.601	0.422	NA	NA	NA	0.510
TSS	0.719±0.046	0.623±0.044	0.312	0.671	0.512	NA	0.539	0.650	0.620

Operational form: results significantly better than Ahmed et al.
Segmented form: results marginally better for TSS

Conclusion

- We used SVM + HMI vector magnetograms through SHARP parameters
- We compared our results mostly with Ahmed et al. (2013): most recent study and amongst the largest database (27539 flaring ARs, 469516 non-flaring ones, incl. C-class)
- Good results overall: TSS is larger than other papers studied, especially in segmented mode
- Probably due to better features: vector magnetograms give access to field topology
- TOTUSJH is the most useful parameter, and only 5-6 are needed: confirms conclusion of Leka and Barnes (2007) that USFLUX, TOTUSJH, and TOTUSJZ are very useful
- However, still far from perfect: 36% (operational mode) and 27% (segmented mode) of flares not predicted
- As was concluded several times in the past by other groups (e.g. Leka and Barnes, 2007), it is not clear that we can improve on flare forecasting with only photospheric magnetograms
- Future works:

take into account time evolution of features, include C-class flares (Colak & Qahwaji, 2009: including C-class flares improve performances), test other ML algorithms (especially SVM+k-nearest neighbors)

AR parameters are sensitive to which pixels contribute to their calculation: should try different masks

Update

- We recently added 2 new features: B effective and fractal dimension: B_{eff} is highly relevant to flaring activity (confirming Georgoulis & Rust, 2007), while the fractal dimension is not relevant (confirming Georgoulis, 2012).

- We started working on including AIA data (discussions with Paul Higgins), in collaboration with Stathis Ilonidis

- We started working on addition of temporal variation in the features (mostly Stathis Ilonidis). Preliminary results show an increase in performance metrics